Quest for early evolution of life based on phylogenetic analyses of aminoacyl tRNA synthetases

Doctoral thesis

Ryutaro Furukawa

Tokyo University of Pharmacy and Life Sciences
Graduate school of Life science
2016

Table of content

1. General Introduction

1.1 Early evolution of life

An ancient history of life before 2 billion years ago is mysterious. Some geological, genetic and biochemical studies have challenged understanding the history of life and proposed numerous hypotheses.

All extant organisms are arisen from one common ancestor. In this thesis, I call the last universal common ancestor thrived at about 4 billion years ago. *Commonote commonote* referred to Akanuma et al. (2015). *C. commonote* diverged to extant organisms in about 4 billion years, but the detailed evolutionary process has been argued. All extant organisms have been classified into three domains (Archaea, Bacteria and Eukarya) by phylogenetic analysis using small subunit ribosomal RNAs (SSU rRNAs) (Woese et al. 1990). Diversification of three domains organisms has been debated between three-domain hypothesis and two-domain hypothesis. In chapter 2 of my thesis, evolutionary history of three domains of life and origin of Eukarya are discussed based on phylogenetic analyses of aminoacyl tRNA synthetase.

We have much less knowledge on evolutionary process of basic biological system before *C. commonote*. Before the origin of life organic compound must have accumulated on the Earth by the process call chemical evolution. Thought the process leading to emergence of life is still an open question. RNA based organisms must have preceded prior to contemporary DNA based organisms: the era of RNA world. Primitive translation system must have appeared in RNA world, before the appearance of *C. commonote*. In chapter 3 of my thesis, evolutionary history of translation system is discussed based on the analyses of aminoacyl tRNA synthetase evolution.

1.2 Aminoacyl tRNA synthetase

Aminoacyl-tRNA synthetases (ARSs) are essential enzymes for translation in all extant organisms. ARSs have been used to resolve early evolution of life because of their universality and sequence conservation (Woese et al. 2000). ARS catalyzes a two-step reaction: 1) The formation of aminoacyl-AMP from amino acid and ATP; and 2) The formation of aminoacyl tRNA from aminoacyl-AMP and tRNA, resulting in the attachment of an amino acid to cognate tRNA. There are more than twenty ARSs and they are classified into two classes, class I and class II, each consisting of three subclasses (a-c) based on the similarity of sequences and structures (Eriani et al. 1990). The classification is the following: class Ia (MetRS, ValRS, LeuRS, IleRS, CysRS and ArgRS), class Ib (GluRS, GlnRS and LysRS-class I), class Ic (TyrRS

and TrpRS), class IIa (SerRS, ThrRS, AlaRS, GlyRS-α_2, ProRS and HisRS), class IIb (AspRS, AsnRS and LysRS-class II), and class IIc (PheRS, GlyRS-$\alpha_2\beta_2$, SepRS and PylRS). In general, ARS consists of a catalytic domain, anticodon-binding domain, and often also an editing domain. Each class harbors class-specific characteristic motifs and structural topology in their catalytic domains (Eriani et al. 1990). Since all known organisms use 20 standard amino acids in translation, the last universal common ancestor is thought to have used the same 20 standard amino acids in translation. There is also the possibility that the diversification of ARSs of each class occurred before the age of last universal common ancestor of all extant organisms (Nagel and Doolittle 1995). The full sets of ARS genes encoded by eukaryal nuclear genomes are classified into cytoplasmic ARS and organellar ARS. No ARS gene is encoded by the organellar genomes. Organellar ARSs are found in either of mitochondria, plastids or apicoplasts. Cytoplasmic ARS is always found in cytosol. In addition, there are “dual-targeted ARSs” that are found in both cytosol and organelles. In this paper, I include the dual-targeted ARSs in cytoplasmic ARSs. Origin and evolution of these enzymes is complex, resulting from various events including gene losses, gene duplications, lateral gene transfers and replacements of other genes, and so on (Wolf et al. 1999; Woese et al. 2000; Brindefalk et al. 2007). Some ARS genes originated from organelles or their ancestral genomes replaced the original cytoplasmic ARS genes during eukaryal evolution (Timmis et al. 2004; Duchêne et al. 2009). Despite the complex evolutionary history, ARS is one of the best genes for the phylogenetic analysis of all extant organisms since the DNA sequences have been well conserved among all domains of life. Therefore, some ARSs were used as core genes for phylogenetic analyses to clarify the relationship between the proposed three domains of life (Wolf et al. 1999; Woese et al. 2000; Brown 2001; 2003).

1.3 Reference

Akanuma S, Yokobori SI, Nakajima Y, Bessho M, Yamagishi A (2015) Robustness of predictions of extremely thermally stable proteins in ancient organisms. Evol 69:2954-2962. doi: 10.1111/evo.12779

Brindefalk B, Viklund J, Larsson D, Thollesson M, Andersson SG (2007) Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. Mol Biol Evol 24:743-756. doi: 10.1093/molbev/msl202

Brown JR (2001) Genomic and phylogenetic perspectives on the evolution of prokaryotes. Syst Biol 50:497-512. doi: 10.1080/10635150117729

Brown JR (2003) Ancient horizontal gene transfer. Nat Rev Genet 4:121-132. doi:10.1038/nrg1000

Duchêne AM, Peeters N, Dietrich A, Cosset A, Small ID, Wintz H (2001) Overlapping destinations for two dual targeted glycyl-tRNA synthetases in *Arabidopsis thaliana and Phaseolus vulgaris*. J Biol Chem 276:15275-15283. doi: 10.1074/jbc.M011525200

Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347:203-206. doi: 10.1038/347203a0

Nagel GM, Doolittle RF (1995) Phylogenetic analysis of the aminoacyl-tRNA synthetases. J Mol Evol 40:487-498. doi: 10.1007/BF00166617

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123-135. doi:10.1038/nrg1271

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579.

Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64: 202-236. doi: 10.1128/MMBR.64.1.202-236.2000

Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 9:689-710. doi: 10.1101/gr.9.8.689

2. Searching for ancestor of Eukarya based on aminoacyl tRNA synthetase

2.1 Background

All extant organisms have been classified into three domains by phylogenetic analysis using SSU rRNAs (Woese et al. 1990). In the three-domain hypothesis, Eukarya is a sister group of Archaea. The three-domain hypothesis has been supported by various molecular phylogenetic studies and phylogenomic studies (Harris et al. 2003; Ciccarelli et al. 2006; Yutin et al. 2008; Rinke et al. 2013).

On the other hand, Lake and coworkers have proposed that some archaeal species are more related to Eukarya than other archaeal species, and suggested that the Eukarya are not an independent domain but located within a group of archaea (Rivera and Lake 1992). The two-domain hypothesis implies that the Eukaryal ancestor was derived from a certain archaeal lineage. The evolutionary relationship of Eukarya and Archaea has been debated between the three-domain hypothesis and the two-domain hypothesis, and several archaeal host hypotheses have been proposed over the last two decades. For example, the eocyte hypothesis describes a close relationship between a crenarchaeotal ancestor and Eukaryota, and has been supported by phylogenetic analysis of SSU rRNA and indel analysis of translational elongation factor (Rivera and Lake 1992). Several phylogenetic analyses using ribosomal proteins, translation factors, and concatenated data of core genes have indicated that TACK superphylum is the most closely related species to Eukarya (Kelly et al. 2011; Guy and Ettema 2011; Williams et al. 2012; Lasek-Nesselquist and Gogarten 2013; Guy et al. 2014; Williams and Embley 2014). Based on the concatenated phylogenetic analyses and comparative genome analyses, Martijn and Ettema also proposed a ‘phagocytosing archaean theory’ (phAT), which describes five steps towards the emergence of eukaryotic cells (Martijn and Ettema 2013).

Methanogen were proposed to be an archaeal ancestor of Eukarya, 18 years ago (Martin and Muller 1998; López-García and Moreira 1999). This hydrogen hypothesis (Martin and Muller 1998) or syntrophic hypothesis (López-García and Moreira 1999), proposed that methanogen and one or more bacteria shared different metabolic sources and an endosymbiotic event occurred gradually in the low nutrient environment. Recently, large-scale single gene phylogenetic analysis showed that euryarchaeotal genes are most frequently placed as a sister to the Eukarya clade (Thiergart et al. 2012). Thiergart et al. also suggested that these analyses supported a methanogenic archaeal host for Eukarya genesis.

Recent innovations in deciphering microbial dark matter and metagenome data

provided information on uncultivated Bacterial and Archaeal genomes (Rinke et al. 2013; Castelle et al. 2015). It potentially improves understanding of the phylogenetic relationships among the three domains. DPANN superphylum consisting of ultra-small cellular archaea (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota, Micrarchaeota) was proposed as a new archaeal group by phylogenetic analysis based on the concatenated protein genes (Rinke et al. 2013). In addition, the genome sequences of Woesearchaeota and Pacearchaeota were reconstructed and then they were classified into DPANN superphylum (Castelle et al. 2015).

One of the recent discoveries on the origin of Eukarya is the discovery of a new archaeal phylum Lokiarchaeota (Spang et al. 2015). The Lokiarchaeota was suggested to be the closest relatives of Eukarya based on the phylogenetic analyses of universally conserved protein genes. The lokiarchaeotal genome was also reported to carry the signature proteins of Eukarya related to cytoskeleton, membrane remodeling and phagocytosis, suggesting that it is an ancestor of Eukarya.

Large-scale single gene phylogenetic analyses using more recent data showed that Eukaryal genes were nested with either TACK superphylum or Euryarchaeota depending on the genes, which hide the true archaeal ancestor of Eukarya (Rochette et al. 2014; Pittis and Gabaldón 2016). These analyses also suggested that many eukaryal genes were nested with several bacterial species, which show that lateral gene transfers from several bacteria lineages contributed to the formation of the last eukaryal common ancestor (LECA) (Thiergart et al. 2012; Rochette et al. 2014; Ku et al. 2015; Pittis and Gabaldón 2016). As proposed at the end of this chapter, I refer to LECA as *Commonote eukaryotes* and also abbreviate this species as *C. eukaryotes.*

In this chapter, I reconstructed and compared the single-gene phylogenetic trees using 23 aminoacyl tRNA synthetases to clarify the phylogenetic relationship among Eukarya, Archaea and Bacteria, by incorporating increased sequence data of various recently discovered organisms. Previous phylogenetic analyses of ARSs supported the three-domain hypothesis (Wolf et al. 1999; Woese et al. 2000; Brindefalk et al. 2007). However, no sequences from TACK superphylum were used in the phylogenetic study by Brindefalk et al. (2007). Thus it is important to conduct a molecular phylogenetic analysis of ARSs that includes new archaeal species and innovative technology to test the three-domain and the two-domain hypotheses. Based on our phylogenetic analyses, we proposed a model for how Eukarya became established.

2.2 Materials and Methods

2.2.1 Sequence Data of ARS

We selected two or three typical species from each order to reduce taxonomic bias. All protein sequences of 282 selected organisms (Archaea: 76, Bacteria: 142, Eukarya: 64) were collected from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). We constructed a KF database (M. Kanetake, R. Furukawa, S. Yokobori, and A. Yamagishi, unpublished) in Geneious ver. 7.1.9 (http://www.geneious.com, Kearse et al. 2012) that consisted of all protein sequences of 282 organisms. The KF database was first constructed on 14 October 2010 and was last updated on 6 June 2015. Protein sequences of 23 ARSs were searched with BlastP (Altschul et al. 1997) from the KF database. Accession numbers of all collected data is shown in Supplemental table S1.

2.2.2 Sequence Alignment

Amino acid sequences of each ARS were aligned using MAFFT 7.017 (Katoh et al. 2013) and edited manually. The editing domain in bacterial LeuRS is located after the ZN-1 domain, whereas the editing domain is located before the ZN-1 domain in archaeal/eukaryal LeuRS, IleRS and ValRS (Cusack et al. 2000). The editing domain in bacterial LeuRS was transferred in front of the ZN-1 domain during the manual alignment. Standard bacterial GlyRS-$\alpha_2\beta_2$ consists of separate α subunit and β subunit genes, while GlyRS-$\alpha_2\beta_2$ in Chlamydia and organelles in plants have fused α-β subunit (Wager et al. 1995; Duchêne et al. 2001). We refer to this concatenated sequence as GlyRS-2 and standard GlyRS distributed in Archaea, Eukarya and some Bacteria as GlyRS-1. The sequences of α and β subunits of GlyRS-2 were concatenated to test the evolutionary relationship between bacteria and organellar GlyRS in plants. The well-aligned regions of each alignment were selected from the final alignment using TrimAl 1.4 (Capella-Gutierrez et al. 2009). TrimAl was used with automated1 mode and the columns containing gap were excluded with nogaps mode. The numbers of sites of the final alignment of 23 ARSs are shown in supplemental table S2.

2.2.3 Phylogenetic analysis

The optimal amino acid substitution model for each ARS alignment was selected using the model selection program PROTTEST 3.4 (Darriba et al. 2011) and is shown in supplemental table S2. We reconstructed trees for 23 ARSs using Maximum likelihood (ML) and Bayesian Inference (BI) analyses. ML analyses were done with the program RAxML 8.1

(Stamatakis 2014) with optimal amino acid substitution model for each ARS. RELL bootstrap analysis was done by analyzing 1000 resampled data sets (Minh et al. 2013). Posterior-probability consensus trees in BI analysis (BI trees) were constructed using PhyloBayes 3.3f (Lartillot et al. 2009) by running two chains until the max discrepancy dropped lower than 0.3 under the CAT Poisson + Γ(4) model. The consensus tree was output using the readpb program. The trees used to readpb analysis were sampled every 10 generations in each analysis. The number of cut-off trees and the reached generation of chains in each analysis are shown in supplemental table S2.

2.2.4 Tree reconstruction of the universal tree based on the small subunit rRNA sequences

The SSU rRNA tree was reconstructed for reference. The initial tree was reconstructed using RAxML with the GTR + Γ model, based on 261 SSU rRNA sequences (Supplemental Fig. S1). SSU rRNA sequences were downloaded from Silva database (Quast et al. 2013) or were directly extracted from genome sequences of each organism, which were downloaded from NCBI (http://www.ncbi.nlm.nih.gov). The root of the tree was placed between Bacteria and Archaea based on previous composite tree analyses (Iwabe et al. 1989, Brown and Doolittle 1995, Zhaxybayeva et al. 2005).

2.3 Result

2.3.1 Phylogenetic reconstruction of 23 ARSs

Phylogenetic trees of 23 ARS genes [AlaRS, ArgRS, AspRS, AsnRS, CysRS, GluRS, GlnRS, GlyRS-1, GlyRS-2, HisRS, IleRS, LeuRS, LysRS-class I, LysRS-class II, MetRS, PheRS-α, PheRS-β, ProRS, SerRS, ThrRS, TrpRS, TyrRS and ValRS] were constructed by using ML and BI analyses (Figs.1-3 and Supplemental Fig. S2). We first checked the eukaryal monophyly in 23 trees. Eukaryal monophyly, with all Eukaryal taxa in one clade, allows tracing back to *C. eukaryotes* and the identification of the closest prokaryotic species to *C. eukaryotes*.

Eukarya generally have cytoplasmic type ARS and organellar type ARS. We evaluated whether cytoplasmic ARSs were a monophyletic group in each tree. Eukaryal monophyly of cytoplasmic ARS was supported with 100% RELL bootstrap support values (rbp) in ML analyses and > 0.99 posterial probability (pp) in BI analyses in 12 ARS trees (SerRS, GlyRS-1, LeuRS, GluRS, TrpRS, PheRS-α, PheRS-β, ValRS, LysRS-class II, ThrRS, IleRS, AspRS) (Figs. 1 and 2 and Supplemental Fig. S2). Eukaryal cytoplasmic ARSs formed a monophyletic ingroup of Archaea in 7 out of 12 trees (SerRS, GlyRS-1, LeuRS, GluRS, TrpRS, PheRS-α, PheRS-β) (Figs. 1 and Supplemental Fig. S2). Eukaryal cytoplasmic ARS was an ingroup of Bacteria in ValRS, LysRS-class II and ThrRS trees (Figs. 2 and Supplemental Fig. S2). Eukaryal cytoplasmic IleRS and AspRS were ingroups of the bacterial group in the archaea group. No monophyletic eukaryal cytoplasmic ARSs were placed as the independent group from bacterial ARSs and archaeal ARSs. Thus, these ARS trees supported the two-domain hypothesis.

On the other hand, eight eukaryal cytoplasmic ARSs were split into two or three groups in the trees of CysRS, AsnRS, TyrRS, ProRS, HisRS, AlaRS, ArgRS, and MetRS (Figs. 3 and Supplemental Fig. S2). In these trees, one cytoplasmic ARS might have originated from that of *C. eukaryotes*, and the others are presumed to have been transferred from prokaryotes through lateral gene transfer during diversification of Eukarya. When the transferred ARS was adapted to the recipient cell, the original cytoplasmic ARS may have disappeared from the Eukaryal genome or may have been maintained for another function.

Eukaryal cytoplasmic ARS was absent in LysRS-class I and GlyRS-2 trees (Figs. 1 and 2 and Supplemental Fig. S2) as reported in preceding studies (Wolf et al. 1999; Woese et al. 2000; Brindefalk et al. 2007). Eukaryal cytoplasmic GlnRS was a sister group of the bacterial GlnRS group. Since GlnRS evolved from eukayral GluRS by gene duplication during the early

evolutionary stage of Eukarya (Lamour et al. 1994; Siatecka 1998; Brown and Doolittle 1999; Woese et al. 2000; Nureki et al. 2010), eukaryal cytoplasmic GlnRS was not derived from bacterial ones; instead, bacterial GlnRS was derived from eukaryal ones by lateral gene transfer (Supplemental Fig. S3).

Organellar ARSs were placed in the bacterial group in most trees, whereas some other organellar ARSs were ingroups of the eukaryal cytoplasmic group. Organellar ARSs in the bacterial group suggested that lateral gene transfer or endosymbiotic gene transfer occured from Bacteria to Eukarya, which may be an important lead to trace back the evolution of organellar ARSs and origin of Eukarya (Brindefalk et al. 2007). Organellar ARSs in the Eukaryal group might have been created by gene duplication during Eukaryal evolution.

2.3.2 Archaeal origin of eukaryal ARSs

Seven eukaryal cytoplasmic ARSs (SerRS, GlyRS-1, LeuRS, GluRS, TrpRS, PheRS-α, PheRS-β) were an ingroup of Archaea, indicating that the seven eukaryal cytoplasmic ARSs were derived from Archaea (Fig.1 and Supplemental Fig. S2). The closest Archaeal taxa to Eukaryal cytoplasmic ARSs are listed in Table 1.

Eukaryal cytoplasmic SerRS was the closest to the monophyletic group consisting of lokiarchaeotal SerRS and *Methanobacterium lacus* (a member of class Methanobacteria of Euryarchaeota) SerRS. Previous phylogenetic analysis of SerRS showed that most methanogenic archaea have a rare-form of SerRS, and these sequences showed little similarity to the common-form SerRS of other Archaea, Bacteria and Eukarya (Kim et al. 1998; Andam and Gogarten 2011). In our study, the rare-form SerRS sequences were removed from the final alignment for our phylogenetic analysis but were listed in the supplemental table S1. Andam and Gorgaten (2011) also proposed that ancient gene duplication occurred before the establishment of last universal common ancestor and ancient SerRS diverged to the rare-form and common-form. The common ancestor of most methanogenic archaea acquired the rare-form SerRS through lateral gene transfer from an extinct lineage and lost the common form SerRS (Andam and Gogarten 2011). However, SerRS of *Methanobacterium lacus* retained the common-form SerRS group. Most methanobacterial species retain rare-form SerRS (supplemental table S1), suggesting only *Methanobacterium lacus* acquired SerRS from Lokiarchaeota through lateral gene transfer very recently. Thus, the closest archaeal species of Eukarya is judged to be Lokiarchaeota in the SerRS tree, suggesting that Eukarya cytoplasm

was derived from a member of TACK superphylum. This relationship is consistent with previous concatenated gene based phylogenetic studies (Guy and Ettema 2011; Kelly et al. 2011; Williams et al. 2012; Lasek-Nesselquist and Gogarten 2013; Guy et al. 2014; Williams and Embley 2014), especially a recent metagenomic analysis that proposed the Lokiarchaeota as the eukaryal ancestor or the closest relative of Eukarya (Spang et al. 2015).

On the other hand, eukaryal cytoplasmic GlyRS-1 was a sister group of Euryarchaeotal GlyRS-1 (Fig.1 and Supplemental Fig. S2). The GlyRS-1 tree indicates that the eukaryal cytoplasm was derived from Euryarchaeota. The gene trees where Euryarchaeota is the closest relative to Eukarya were observed in previous large-scale single gene studies (Thiergart et al. 2012; Rochette et al. 2014; Ku et al. 2015; Pittis and Gabaldón 2016).

Furthermore, the closest species of monophyletic cytoplasmic ARSs in three trees (GluRS, LeuRS, TrpRS) were certain species of DPANN superphylum. These results suggest that the ancestor of Eukarya was the DPANN superphylum of Archaea. However, the closest archaeal phyla of each eukaryal ARS were different (GluRS: Micrarchaeota, LeuRS: Parvarchaeota, TrpRS: Woesearchaeota). The second closest species of eukaryal GluRS was thaumarchaeotal GluRS, the second closest of eukaryal LeuRS was LeuRS from Crenarchaeota, Aigarchaeota, and several DPANN archaea and the second closest of eukaryal TrpRS was TrpRSs from the group of several TACK archaea, Thermococci and several DPANN archaea, which suggest the possibility that the true ancestor of 3 eukaryal cytoplasmic ARSs (GluRS, LeuRS, and TrpRS) were those of TACK superphylum and the single sister DPANN archaea may be the result of gene transfer from the ancestor of TACK superphylum.

The closest species of monophyletic cytoplasmic ARSs in five trees (SerRS, GlyRS-1, GluRS, LeuRS, TrpRS) showed that Eukarya derived from three Archaea groups: TACK superphylum, Euryarchaeota and DPANN superphylum. In either case, our results were different from previous ARS phylogenetic studies that support the three-domain hypothesis (Wolf et al. 1999; Woese et al. 2000; Brindefalk et al. 2007). In their analyses, only limited archaeal species were included (Supplemental table S3). Thus, our results show a more detailed phylogenetic relationship between archaeal phyla and support the two-domain hypothesis instead of the three-domain hypothesis with abundant taxon sampling. Abundant taxon sampling and optimal evolutionary models provide more accurate evolutionary relationships.

In two cytoplasmic ARS trees (PheRS-α, PheRS-β), the identification of an archaeal ancestor of Eukarya was difficult because the closest group of two cytoplasmic ARSs contained

several species of archaeal phyla. However, these trees also support the two-domain hypothesis. The closest species of PheRS-α was the group of several Euryarchaeota and several DPANN archaea and the closest species PheRS-β was the group of Euryarchaeota and most TACK archaea. Though PheRS is heterotetramer enzyme consising of two PheRS-α subunits and two PheRS-β subunits, which imply that the two genes should trace the same evolution, these phylogenetic histories are different within the archaeal lineage. Previous genome analyses showed that most archaeal PheRS-α and PheRS-β were encoded on a different operon from each other (Brown 2001), which suggests that the two subunits evolved independently. Thus, the difference between the two trees is the result of independent evolution of PheRS-α and PheRS-β associated with lateral gene transfer between archaeal species.

2.3.3 Bacterial origin of eukaryal cytoplasmic ARSs

Monophyly of eukaryal cytoplasmic ARSs derived from bacterial ones was found in five trees (ValRS, ThrRS, IleRS, AspRS, LysRS-class II) (Fig. 2 and S1). The closest species of eukaryal cytoplasmic ARSs are shown in Table 2. ValRS tree suggested that eukaryal cytoplasmic ValRS derived from *Myxococcus xanthus* supported by 89% rbp in ML analyses and 0.99 pp in BI analyses. The sister group of eukaryal cytoplasmic ThrRS consists of three bacterial phyla (Gemmatimonadetes, Deltaproteobacteria (*Myxococcus xanthu*), Poribacteria). Eukarya cytoplasmic IleRS and AspRS derived from the bacteria group in Archaea, which shows that some bacteria acquired archaeal genes to adapt to the environment at least once through lateral gene transfer and *C. eukaryotes* acquired the archaeal gene from Bacteria (Brown et al. 2003). Eukaryal cytoplasmic IleRS is the sister group of Lentisphaera. Eukaryal cytoplasmic AspRS is the sister group of some bacterial phyla (ML tree: Deinococcus-Thermus, Spirocheta, *Candidatus Acetothermus*, *Clostridium*, Microgenomates, BI tree: Candidate division WWE3, Candidate division WS6, Peregrinibacteria). However, the phylogenetic position of eukaryal cytoplasmic LysRS-class II was difficult to determine because the closet species to Eukarya was different between ML and BI trees. Eukaryal cytoplasmic LysRS-class II was the sister group of Archaea in the ML tree, but cytoplasmic LysRS-class II was the closest to Aquificae in the BI tree.

Four monophyletic eukaryal cytoplasmic ARSs (ValRS, ThrRS, IleRS, and AspRS) were closest to various bacterial species (Fig. 3 and Supplemental Fig. S2), suggesting that independent lateral gene transfer occurred from the bacterial genome to the genome of *C.*

eukaryotes and replaced the cytoplasm ARS. Various bacterial lateral gene transfers in our phylogenetic trees supported the slow-drip hypothesis (Rochette et al. 2014), which proposed that the stem eukaryotic ancestor acquired bacteria-related eukaryotic genes through lateral gene transfer from mitochondria-unrelated Bacteria. Similar bacterial gene transfers were observed in previous studies, whose genes mainly contribute metabolic function (Yutin et al. 2008; Saruhashi et al. 2008; Thiergart et al. 2012, Ku et al. 2015). Recent single gene tree analysis shows that gene transfers from various bacteria contributed to eukaryogenesis before endosymbiosis of α-proteobacteria (Pittis and Gabaldón 2016).

2.3.4 Origin of cytoplasmic ARS in the polyphyletic Eukarya tree

Eight eukarya cytoplasmic ARSs (CysRS, AsnRS, TyrRS, ProRS, HisRS, AlaRS, ArgRS, MetRS) were split into 2 or 3 groups in each phylogenetic tree. These trees showed that after the Eukarya acquired cytoplasmic ARS, some eukaryal species acquired another cognate ARS through lateral gene transfer or endosymbiotic gene transfer and the original ARS may have been lost. Alternatively, *C. eukaryotes* have had 2 or 3 genes of each ARS and differential genes were lost in each eukaryal lineage. Comparing two theories, since each eukaryal species has only one set of cytoplasmic eukaryal ARS, the acquisition of foreign genes after the divergence of Eukarya is parsimonious and reasonable. Thus, we needed to estimate which ARS is original and which is secondary in individual trees. Phylogenetic trees and the closest species are shown in Fig. 3, S1 and Table 3, respectively.

In four trees (CysRS, AsnRS, TyrRS ProRS), one cytoplasmic ARS was derived from Archaea and the other was derived from Bacteria or another Archaeal group, indicating that Eukaryal cytoplasmic ARS were derived from Archaea first and then Eukarya acquired Bacterial or Archaeal ARSs during Eukaryal evolution or that *C. eukaryotes* acquired secondary ARS and differential loss of ARS occurred in each Eukaryal lineage later.

In the CysRS tree, eukaryal cytoplasmic CysRS, with the exception of some plants, was the sister group of *Methanococcus* and Thermococci, indicating that most eukaryal cytoplasmic CysRS derived from Euryarchaeota. Cytoplasmic CysRS of some plants and organellar CysRS of some plants were derived from proteobacteria, suggesting lateral gene transfer from proteobacteria to the plants. Then the plant organellar CysRS would have duplicated and one of the two organellar CysRSs replaced the cytoplasmic CysRS during evolution of these plants.

In the AsnRS tree, the cytoplasmic AsnRS of Excavata, Metazoa, Fungi, and Amoebozoa formed a monophyletic group as the ingroup of Archaea and the sister group was the phylum Micrarchaeota, a member of DPANN superphylum. AsnRS of Plants, Stramenopiles and Alveolata were ingroups of Bacteria, but the sister group could not be clarified in both ML and BI trees because the taxon of the first eukaryal group consists of a wide range of taxa. The ancestor of eukaryal cytoplasmic AsnRS may be closely related to Micrarchaeota. Endosymbiotic gene transfer of organellar AsnRS may have occurred in the common ancestor of Plants, Stramenopiles and Alveolata. Monophyletic relationship of Plants, Stranmenopiles and Alveolata was recovered in some phylogenomic analyses (Philippe et al. 2004; Simpson et al. 2006; Burki et al. 2008, 2009; Derelle and Lang 2012; Zhao et al. 2012; Katz and Grant 2015; Cavalier-Smith et al. 2015; Kamkowska et al. 2016).

In two trees (TyrRS, ProRS), one cytoplasmic ARS branch was placed in one Archaea group and the other placed in a different Archaea group. A clade of eukaryal cytoplasmic TyrRS, except Metazoa, Fungi and Acanthamoeba, were the sister groups of woesearchaeotal TyrRS. Since TyrRS in a wide range of eukaryal taxa was derived from Woesearchaeota, TyrRS of *C. eukaryotes* originated from DPANN superphylum. The common ancestor of Metazoa, Fungi, and Acanthamoeba acquired another archaeal TyrRS of DPANN superphylum before diversification of Metazoa, Fungi, and Acanthamoeba. In the ProRS tree, most Eukaryal cytoplasmic ProRSs formed a sister group of Woesearchaeota and the other cytoplasmic ProRSs of a few excavates formed a sister group of Aigarchaeota. Thus, the *C. eukaryotes* possessed ProRS acquired from the closely related organisms of Woesearchaeota. A few species of excavates acquired ProRS from the closely related organisms of Aigarchaeota through lateral gene transfer and lost the ProRS from the closely related organisms of Woesearchaeota.

In the ML tree of HisRS, most eukaryal cytoplasmic HisRSs were sister groups of various Archaea, especially TACK superphylum in the ML tree. However, the group appeared as the sister groups of Peregrinibacteria in the BI tree. Accordingly, the ancestor of cytoplasmic HisRS is still unclear. Remaining eukaryal cytoplasmic HisRSs (those of Euglenozoa, Algae, Stramenopiles, Naegleria and Acanthamoeba) formed a sister group of various Bacteria, which shows that their HisRS derived from Bacteria through lateral gene transfer.

In the AlaRS tree, the eukaryal clade consisting of Metazoa, Fungi, Amoebozoa except for Entamoeba, Plants, Alveolata except for Ciliophora, Stramenopiles, Cryptophyta, Heterolobosea and Euglenozoa was an ingroup of Bacteria and was the sister group of

Phycisphaeria. On the other hand, the eukaryal clade consisting of fewer taxonomic groups including Diplomonadida, Trichomonadida, Ciliophora and Entamoeba was an ingroup of Archaea and was the sister group of various Archaeal groups. AlaRS indicated that *C. eukaryotes* acquired AlaRS from Phycisphaeria and that secondary lateral gene transfer occurred from archaeal species to fewer taxonomic eukaryal groups during eukaryal evolution. Then AlaRS of Phycisphaeria was adopted as cytoplasmic and mitochondrial ARS in the translation system of *C. eukaryotes*. This result is consistent with previous AlaRS analysis, which showed that most eukaryal AlaRSs formed an ingroup of Bacteria and that those of Diplomonadida, Parabasalia, Ciliophora and Entamoeba formed sister groups of nanoarchaeote AlaRS (Andersson et al. 2005). These reports suggested that lateral gene transfer occurred from Nanoarchaota to the common ancestor of Diplomonadida and Parabasalia first, and then lateral gene transfer occurred from Diplomonadida or Parabasalia to Ciliophora and Entamoeba (Andersson et al. 2005).

Eukaryal cytoplasmic ArgRS and MetRS derived from bacterial ones through three independent lateral gene transfer events during evolution of Eukarya. In the ArgRS tree, Eukarya except Fungi, Amoebozoa and red algae was a sister group of Chlamydiae. Cytoplasmic ArgRS of Fungi and Amoebozoa and organellar ArgRS of Metazoa were a sister group of *Myxococcus*. Cytoplasmic ArgRS of red algae was the sister group of Cyanobacteria. Summarizing these results, *C. eukaryotes* acquired ArgRS of Chlamydiae first. Second, the common ancestor of Fungi, Amoebozoa and Metazoa acquired ArgRS from *Myxococcus* as the mitochondrial ArgRS, and third, cytoplasmic ArgRS of Fungi and Amoebozoa was replaced by mitochondrial ones. The common ancestor of red algae acquired ArgRS from Cyanobacteria through each independent gene transfer.

In the MetRS tree, cytoplasmic MetRSs of Metazoa, Fungi, Plants, Amoebozoa and a part of Alveolata formed a monophyletic ingroup of Spirochete MetRSs. Cytoplasmic MetRSs of Euglenozoa, Excavata and organellar MetRSs of Metazoa and Fungi formed a monophyletic ingroup of Candidate division TM6 with 92% rbp in ML analyses and 0.98 pp in BI analyses. Cytoplasmic MetRS of most Alveolata, Stramenopiles and green algae were also placed in the Bacterial group. Since the majority of cytoplasmic MetRS were derived from Spirochetes, *C. eukaryotes* acquired MetRS from Spirochetes first. Two independent gene transfer events from a bacterial ancestor occurred after gene transfer of Spirochetes and the transferred MetRS replaced the cytoplasmic MetRS in some eukaryal taxa during evolution.

Eight polyphyletic cytoplasmic ARSs showed that independent lateral gene transfer from Archaea or Bacteria occurred during evolution of Eukarya and the transferred genes replaced the cytoplasmic ARS genes. *C. eukaryotes* might have four ARSs of archaeal origin (CysRS, AsnRS, ProRS and TyrRS), three ARSs of bacterial origin (AlaRS, ArgRS and MetRS) and 1 HisRS of unknown origin. Specifically, 1 Archaeal ARS (CysRS) derived from Euryarchaeota and 3 Archaeal ARSs (AsnRS, ProRS and TyrRS) derived from DPANN superphylum. These could be explained with an alternative possibility; *C. eukaryotes* may have had 2 genes of each ARS and differential genes were lost in each Eukaryal lineage. Recent single gene phylogenetic analysis also proposed that patchy distribution of eukaryal genes is mainly the result of differential gene loss and lateral gene transfer provided a few contributions to evolution of Eukarya (Ku et al. 2015). In any case, ARS from Euryarchaeota, DPANN superphylum and Bacteria have contributed to the evolution of eukaryal cells.

2.4 Discussion

2.4.1 Chimeric origin of eukaryal cells

Since Eukarya have a mosaic genome consisting of Bacterial genes, Archaeal genes and Eukarya specific genes, the origin of Eukarya is one of the most challenging problems in biology. Various fusion models of eukaryal origin were proposed for explaining the mosaic eukaryal genome (Zillig 1991; Martin and Muller 1998; López-García and Moreira 1999; Rivera and Lake 2004; Forterre 2011). Our ARS trees support the theory that the ancestral eukaryal genome was a chimera of genes of bacterial and archaeal origins.

In our ARS study presented here, we observed that 11 eukaryal cytoplasmic ARSs were derived from Archaea and 7 eukaryal cytoplasmic ARSs were derived from Bacteria, whereas no eukaryal cytoplasmic ARSs formed a third group independent from bacterial and archaeal counterparts. These observations do not fit with the three-domain hypothesis proposed by Woese et al. (1990). Among 11 ARS trees in which eukaryal ones appeared as the ingroup of archaeal ARSs, only one ARS (SerRS) was compatible with the hypothesis of TACK superphylum as the eukaryal ancestor. The phylogenetic analyses of selected concatenated genes supported the TACK superphylum as an ancestor of Eukarya (Guy and Ettema 2011; Kelly et al. 2011; Williams et al. 2012; Lasek-Nesselquist and Gogarten 2013; Guy et al. 2014; Williams and Embley 2014; Spang et al. 2015). Also, single gene phylogenetic analyses of 5 highly conserved proteins using concatenated genes phylogenetic analysis supported Lokiarchaeota as the closest to Eukarya, although the other single gene trees of 30 proteins using concatenated genes phylogenetic analysis show low resolution at the critical node between archaea and Eukarya (Spang et al. 2015). These studies supported a closer relationship between Eukarya and Lokiarchaeota. Our analysis on SerRS also supported this relationship. However, considering the low resolution between Lokiarchaeota and other phyla of TACK superphylum in our SerRS tree, we cannot judge which phylum of TACK superphylum, including Lokiarchaeota is closest to Eukarya. We conclude that Eukarya has their origin within TACK superphylum based on the phylogenetic analysis of SerRS.

However, our BlastP analysis did not detect ValRS and TyrRS in Lokiarchaeota as shown in Supplemental Table S1. In addition, only incomplete sequence of MetRS gene of Lokiarchaeota was detected by our BlastP analysis. These results imply that incomplete genome sequence of lokiarchaeota makes it difficult to detect these ARSs or genome reduction may have occurred in the Lokiarchaeota lineage specifically. Thus, further analyses are desired by

using a more complete genome of Lokiarchaeota.

Moreover, a close relationship between Euryarchaeota and Eukarya was also observed in our analysis (GlyRS-1, CysRS), and was reported in previous studies (Thiergart et al. 2012; Rochette et al. 2014, Pittis and Gabaldón 2016). These relationships support a euryarchaeotal ancestor of Eukarya, as proposed by the hydrogen hypothesis (Martin and Muller 1998) and the syntrophy hypothesis (Moreira and Lopez-garcia 1998). Since euryarchaeotal ancestry of eukaryotic genes is not a minor case in single gene phylogenetic analyses (Thiergart et al. 2012; Rochette et al. 2014, Pittis and Gabaldón 2016), we cannot ignore the contribution of Euryarchaeota to the formation and evolution of eukaryotic cell. Perhaps there was frequent lateral gene transfer from Euryarchaeota to the archaeal ancestor of Eukarya.

The third ancestor related to DPANN superphylum is the closest relative to Eukarya, and was observed in 6 ARS trees (GluRS, LeuRS, TrpRS, TyrRS, AsnRS and ProRS). DPANN superphylum was a monophylic group and was far from the Eukarya group in the concatenated phylogenetic analyses (Rinke et al. 2013; Willams and Embley 2014). Recent phylogenetic analysis classified Woesearchaeota and Pacearchaeota as members of DPANN superphylum (Castelle et al. 2015). Since analyzed species of DPANN superphylum have a small genome that has lost genes of some enzymes for metabolism, it is suggested that the lifestyle of species belonging to DPANN superphylum are symbiotic or parasitic (Castelle et al. 2015). In previous concatenated protein phylogenetic trees, the phylogenetic position of DPANN superphylum is far from Eukarya (Williams and Embley 2014; Spang et al. 2015; Castelle et al. 2015). In our analyses of 6 ARSs, Parvarchaeota, Micrarchaeota and Woesearchaeota were closer species to Eukarya than other phyla of DPANN superphylum. These relationships suggested a symbiotic or parasitic life style between these DPANN taxa (Parvarchaeota, Micrarchaeota, and Woesearchaeota), which we call the PMW group and *C. eukaryotes*. However, a monophyletic group of DPANN superphylum or PMW group never appears in our trees, which suggests that DPANN superphylum is an unreliable classification of archaeal phylum. Symbiotic gene transfers were observed between *Ignicoccus hospitalis* and *Nanoarchaeum equitans* (Rachel et al. 2002; Podar et al. 2008), which suggest that independent gene transfers from each symbiotic archaeal species is realistic. A symbiotic relationship might have occurred by independent gene transfers from each PMW taxa to the ancestor of Eukarya. Thus, gene transfers from each PMW taxa obviously contributed to the evolution of Eukarya. These gene transfers were hidden in

previous single phylogenomic studies because these analyses contained few species of DPANN superphylum (Thiergart et al. 2012; Rochette et al. 2014; Ku et al. 2015; Pittis and Gabaldón 2016).

Bacterial species as eukaryotic ancestors are consistent with previous single phylogenomic studies (Esser et al. 2004; Thiergart et al. 2012; Rochette et al. 2014; Pittis and Gabaldón 2016). *C. eukaryotes* acquired bacterial genes for energy production through endosymbiotic gene transfer or lateral gene transfer. A recent study provided evidence that some independent lateral gene transfer from various bacterial groups obviously occurred before the endosymbiotic event of α-proteobacteria and promoted the evolution of proto-eukaryal cells (Pittis and Gabaldón 2016). Our ARS trees (ThrRS, ValRS, IleRS, AspRS, AlaRS, ArgRS and MetRS) of bacterial ancestry are consistent with non α-proteobacterial gene transfer before an endosymbiotic event and acquisition of bacterial ARS that might have contributed to adaption of the transferred bacterial tRNA genes.

Summarizing our ARS analyses, *C. eukaryotes* probably had genes of TACK superphylum, Euryarchaeota, DPANN superphylum and some Bacteria. Explaining these complex gene ancestries of Eukarya, Koonin and Yutin (2014) suggested that either the archaeal ancestor of eukarya arose from genome streamlining or was not derived from any direct archaeal lineage (Koonin and Yutin 2014). Our results cannot disprove the theory of Koonin and Yutin, but more phylogenetic analyses using the genes of DPANN superphylum may resolve the complexity of origin of Eukarya.

On the other hand, our ARS analyses tend to be congruent with recent single gene phylogenomic analysis (Pittis and Gabaldón 2016) that detected the chimeric origins of *C. eukaryotes* genes. They also inferred the evolutionary scheme of *C. eukaryotes* genes by measuring the stem length between the eukaryal root point and divergent point against the sister group of Eukarya. The stem lengths of archaeal genes tended to be longer than bacterial genes, which shows that archaeal genes of eukaryal cells are ancient and the eukaryal root arose from Archaea, and also supports the two-domain hypothesis. Both eukaryal genes originated from TACK superphylum and eukaryal genes originated from Euryarchaeota were detected with a similar number of genes in their analysis. However, the stem length of 30 genes that originated from Lokiarchaeota tended to be shorter than that of genes originating from other archaea (Extended Data Figure 7 in Pittis and Gabaldón 2016), which supports Lokiarchaeota as the closest species of Eukarya. Thus, this single gene phylogenetic analysis implies Lokiarchaeota

as the closest origin to *C. eukaryotes* (Spang et al. 2015).

2.5 Proposal

Eukaryal cytoplasmic ARSs were ingroups of Archaea or ingroups of Bacteria ARS in our ARS analysis, which conflicts with the three-domain hypothesis. The eukaryal cytoplasmic ARS set has a chimeric origin. Each ARS tree seems to be consistent with the two-domain hypothesis, although origin of five eukaryal cytoplasmic ARSs are Bacteria rather than Archaea. Based on these observations and our discussion above, we propose a new description on the high-level taxonomy of life (Fig. 4). This model is shown as the tree structure that is based on the SSU rRNA tree constructed by the ML method. Without any changes of topology within each domain, the position of Eukarya can be moved next to Lokiarchaeota (Fig. 4).

In our proposal, we accept that the cytoplasm of Eukarya originated from Lokiarchaeota (or TACK superphylum) (Spang et al. 2015). Then, the “proto-eukaryotic” cells accepted genes from Euryarchaeota, DPANN superphylum, and Bacteria except for α-proteobacteria via lateral gene transfer events. In particular, we emphasize the important contribution of DPANN superphylum for the eukaryogenesis, as we discovered numerous lateral gene transfer events from DPANN superphylum to *C. eukaryotes*. Acquisitions of mitochondria of α-proteobacteria origin (and plastids of cyanobacterial origin) are thought to have followed. The evolutionary scheme of Archaea to *C. Eukaryotes* was proposed in Fig. 5.

Our proposed model (Fig. 4) reflects the main evolutionary history from the last common ancestor of all extant cellular organisms *Commonote commonote* (Akanuma et al. 2015) at about 3.8 billion years ago. The tree of life in our model is divided into Domain Archaea and Domain Bacteria (Table 4). Although the terms “Archaea” and “Bacteria” are taken from Woese et al. (1990), the definitions of them are different. In our definition, Domain Archaea consists of Archaea and Eukarya in Woese et al. (1990). These concepts are referred to previous dichotomic division of the phylogenetic tree of life (Yamagishi and Oshima 1995). We place Archaea and Eukarya within Domain Archaea as Subdomains Archaebacteria and Eukaryotes. Furthermore we propose to define the last eukaryal common ancestor as a species, naming it *Commonote eukaryotes* and also abbreviate this species as *C.eukaryotes*. This naming concept is referred to Akanuma et al. (2015). We assume that *C.eukaryotes* is located at the root position of the Eukaryotic tree. Our and other analyses infer that the proteome of *C.eukaryotes* originated from diverse bacterial and archaeal species (Thiergart et al. 2012; Rochette et al. 2014; Pittis and Gabaldón 2016), which suggests that *C.eukaryotes* had a chimeric genome of bacterial and archaeal origins. Since all extant Eukaryotes have mitochondrial-like organelles

(Gray 2012) except for one Eukaryote (Karnkowska et al. 2016), *C.eukaryotes* would have already acquired mitochondria. We also assume that *C.eukaryotes* is the species that experienced rampant gene transfers from bacteria and archaea, and endosymbiotic events with α-proteobacteria. Domain Bacteria in our definition is not equal to the “Bacteria” in Woese et al. (1990). The “Bacteria” in Woese et al. (1990) is moved to the rank subdomain, and we propose to use “Eubacteria” for the name of this subdomain. In our definition the domain Bacteria consists of Subdomain Eubacteria and eukaryotic organelles with their own genetic system (mitochondria and plastids), although the eukaryotic organelles are not independent cells.

2.6 Reference

Akanuma S, Yokobori SI, Nakajima Y, Bessho M, Yamagishi A (2015) Robustness of predictions of extremely thermally stable proteins in ancient organisms. Evol 69:2954-2962. doi: 10.1111/evo.12779

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402. doi: 10.1093/nar/25.17.3389

Andam CP, Gogarten JP (2011) Biased gene transfer in microbial evolution. Nat Rev Microbiol 9:543-555. doi: 10.1038/nrmicro2593

Andersson JO, Sarchfield SW, Roger AJ (2005) Gene transfers from Nanoarchaeota to an ancestor of diplomonads and parabasalids. Mol Bio Evol 22:85-90. doi: 10.1093/molbev/msh254

Brindefalk B, Viklund J, Larsson D, Thollesson M, Andersson SG (2007) Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. Mol Biol Evol 24:743-756. doi: 10.1093/molbev/msl202

Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci USA 92:2441-2445. doi: 10.1073/pnas.92.7.2441

Brown JR, Doolittle WF (1999) Gene descent, duplication, and horizontal transfer in the evolution of glutamyl-and glutaminyl-tRNA synthetases. J Mol Evol 49:485-495. doi: 10.1007/PL00006571

Brown JR (2001) Genomic and phylogenetic perspectives on the evolution of prokaryotes. Syst Biol 50:497-512. doi: 10.1080/10635150117729

Brown JR, Gentry D, Becker JA, Ingraham K, Holmes DJ, Stanhope MJ (2003) Horizontal transfer of drug - resistant aminoacyl - transfer - RNA synthetases of anthrax and Gram - positive pathogens. EMBO Rep 4:692-698. doi: 10.1038/sj.embor.embor881

Brown JR (2003) Ancient horizontal gene transfer. Nat Rev Genet 4:121-132. doi:10.1038/nrg1000

Burki F, Shalchian-Tabrizi K, Pawlowski J (2008) Phylogenomics reveals a new ‘megagroup’including most photosynthetic eukaryotes. Biol lett 4:366-369. doi: 10.1098/rsbl.2008.0224

Burki F, Inagaki Y, Bråte J et al (2009) Large-scale phylogenomic analyses reveal that two

enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. Genome Biol Evol 1:231-238. doi: 10.1093/gbe/evp022

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-1973. doi: 10.1093/bioinformatics/btp348

Castelle CJ, Wrighton KC, Thomas BC et al (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol 25:690-701. doi: 10.1016/j.cub.2015.01.014

Cavalier-Smith T, Chao EE, Lewis R (2015) Multiple origins of Heliozoa from flagellate ancestors: new cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. Mol phylogenet evol 93:331-362. doi: 10.1016/j.ympev.2015.07.004

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283-1287. doi: 10.1126/science.1123061

Cusack S, Yaremchuk A, Tukalo M (2000) The 2 Å crystal structure of leucyl - tRNA synthetase and its complex with a leucyl - adenylate analogue. EMBO J 19:2351-2361. doi: 10.1093/emboj/19.10.2351

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164-1165. doi: 10.1093/bioinformatics/btr088

Derelle R, Lang, BF (2012) Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol biol evol 29:1277-1289. doi: 10.1093/molbev/msr295

Duchêne AM, Peeters N, Dietrich A, Cosset A, Small ID, Wintz H (2001) Overlapping destinations for two dual targeted glycyl-tRNA synthetases in *Arabidopsis thaliana and Phaseolus vulgaris*. J Biol Chem 276:15275-15283. doi: 10.1074/jbc.M011525200

Duchêne AM, Pujol C, Maréchal-Drouard L (2009) Import of tRNAs and aminoacyl-tRNA synthetases into mitochondria. Curr Genet 55:1-18. doi: 10.1007/s00294-008-0223-9

Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347:203-206. doi: 10.1038/347203a0

Esser C, Ahmadinejad N, Wiegand C et al (2004) A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol

Biol Evol 21:1643-1660. doi: 10.1093/molbev/msh160

Forterre P (2011) A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. Res Microbiol 162:77-91. doi: 10.1016/j.resmic.2010.10.005

Gray MW (2012) Mitochondrial evolution. Cold Spring Harbor perspectives in biology, 4(9), a011403. doi: 10.1101/cshperspect.a011403

Guy L, Ettema TJG (2011) The archaeal 'TACK'superphylum and the origin of eukaryotes. Trends Microbial. 19:580-587. doi: 10.1016/j.tim.2011.09.002

Guy L, Saw JH, Ettema TJG (2014) The archaeal legacy of eukaryotes: a phylogenomic perspective. Cold Spring Harb Perspect Biol 6:a016022. doi: 10.1101/cshperspect.a016022

Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome research 13:407-412. doi: 10.1101/gr.652803

Iwabe N, Kuma KI, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355-9359.

Karnkowska A, Vacek V, Zubáčová Z et al (2016) A eukaryote without a mitochondrial organelle. Curr Biol 26:1274-1284. doi: 10.1016/j.cub.2016.03.053

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780. doi: 10.1093/molbev/mst010

Katz LA, Grant JR (2015) Taxon-Rich Phylogenomic Analyses Resolve the Eukaryotic Tree of Life and Reveal the Power of Subsampling by Sites. Syst Biol 64:406-415. doi: 10.1093/sysbio/syu126

Kearse M, Moir R, Wilson A et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649. doi: 10.1093/bioinformatics/bts199

Kelly S, Wickstead B, Gull K (2011) Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. Proc R Soc Lond B Biol Sci 278:1009-1018. doi: 10.1098/rspb.2010.1427

Kim HS, Vothknecht UC, Hedderich R, Celic I, Söll D (1998) Sequence divergence of seryl-tRNA synthetases in archaea. J Bacteriol 180:6446-6449.

Koonin EV, Yutin N (2014) The dispersed archaeal eukaryome and the complex archaeal

ancestor of eukaryotes. Cold Spring Harb Perspect Biol 6:a016188. doi: 10.1101/cshperspect.a016188

Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-Covo E, Mclnerney JO, Landan G, Martin WF (2015) Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524:427-432. doi: 10.1038/nature14963

Lamour V, Quevillon S, Diriong S, N'guyen VC, Lipinski M, Mirande MARC (1994) Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. Proc Natl Acad Sci U S A 91:8670-8674. doi: 10.1073/pnas.91.18.8670

Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288. doi: 10.1093/bioinformatics/btp368

Lasek-Nesselquist E, Gogarten JP (2013) The effects of model choice and mitigating bias on the ribosomal tree of life. Mol Phylogenet Evol 69:17-38. doi: 10.1016/j.ympev.2013.05.006

López-García P, Moreira D (1999) Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci 24:88-93. doi: 10.1016/S0968-0004(98)01342-5

Martijn J, Ettema TJ (2013) From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem Soc Trans 41:451-457. doi: 10.1042/BST20120292

Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37-41. doi: 10.1038/32096

Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188-1195. doi: 10.1093/molbev/mst024

Nagel GM, Doolittle RF (1995) Phylogenetic analysis of the aminoacyl-tRNA synthetases. J Mol Evol 40:487-498. doi: 10.1007/BF00166617

Nureki O, O'Donoghue P, Watanabe N, Ohmori A, Oshikane H, Araiso Y, Sheppard K, Soll D, Ishitani R (2010) Structure of an archaeal non-discriminating glutamyl-tRNA synthetase: a missing link in the evolution of Gln-tRNA^{Gln} formation. Nucleic acids Res 38:7286-7297. doi: 10.1093/nar/gkq605

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol biol evol 21:1740-1752. doi: 10.1093/molbev/msh182

Pittis AA, Gabaldón T (2016) Late acquisition of mitochondria by a host with chimaeric

prokaryotic ancestry. Nature 531:101-104. doi:10.1038/nature16941

Podar M, Anderson I, Makarova KS et al (2008) A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. Genome biol 9:1-18. doi: 10.1186/gb-2008-9-11-r158

Quast C, Pruesse E, Yilmaz P et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids Res 41:D590-D596. doi: 10.1093/nar/gks1219

Rachel R, Wyschkony I, Riehl S, Huber H (2002) The ultrastructure of Ignicoccus: evidence for a novel outer membrane and for intracellular vesicle budding in an archaeon. Archaea 1:9-18. doi: 10.1155/2002/307480

Rinke C, Schwientek P, Sczyrba A et al (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431-437. doi:10.1038/nature12352

Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74-76. doi: 10.1126/science.1621096

Rivera MC, Lake JA (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431:152-155. doi:10.1038/nature02848

Rochette NC, Brochier-Armanet C, Gouy M (2014) Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. Mol Biol Evol 31:832-845. doi: 10.1093/molbev/mst272

Saruhashi S, Hamada K, Miyata D, Horiike T, Shinozawa T (2008) Comprehensive analysis of the origin of eukaryotic genomes. Genes genet syst 83:285-291. doi: 10.1266/ggs.83.285

Siatecka M, Rozek M, Barciszewski J, Mirande M (1998) Modular evolution of the Glx - tRNA synthetase family. Eur J Biochem 256:80-87. doi: 10.1046/j.1432-1327.1998.2560080.x

Simpson AG, Inagaki Y, Roger AJ (2006) Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of “primitive” eukaryotes. Mol Biol Evol 23, 615-625. doi: 10.1093/molbev/msj068

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJ (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173-179. doi:10.1038/nature14447

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313. doi: 10.1093/bioinformatics/btu033

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF (2012) An evolutionary network of

genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol 4:466-485. doi: 10.1093/gbe/evs018

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123-135. doi:10.1038/nrg1271

Wagar EA, Giese MJ, Yasin B, Pang M (1995) The glycyl-tRNA synthetase of *Chlamydia trachomatis*. J Bacteriol 177:5179-5185.

Williams TA, Embley TM (2014) Archaeal "Dark Matter" and the Origin of Eukaryotes. Genome Biol Evol 6:474. doi: 10.1093/gbe/evu031

Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. Proc R Soc Lond B Biol Sci 279:4870-4879. doi: 10.1098/rspb.2012.1795

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences 87:4576-4579. doi: 10.1073/pnas.87.12.4576

Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64: 202-236. doi: 10.1128/MMBR.64.1.202-236.2000

Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 9:689-710. doi: 10.1101/gr.9.8.689

Yamagishi A, Oshima T (1995) Return to dichotomy: Bacteria and Archaea. Chemical evolution: self-organization of the macromolecules of life. A. Deepak, Hampton, VA, pp 155-158

Yutin N, Makarova KS, Mekhedov SL, Wolf, YI, Koonin, EV (2008) The deep archaeal roots of eukaryotes. Mol Biol Evol 25:1619-1630. doi: 10.1093/molbev/msn108

Zhao S, Burki F, Bråte J, Keeling P, Klaveness D, Shalchian-Tabrizi K (2012) Collodictyon–an ancient lineage in the tree of eukaryotes. Mol biol evol 29:1557-1568. doi: 10.1093/molbev/mss001

Zhaxybayeva O, Lapierre P, Gogarten JP (2005) Ancient gene duplications and the root (s) of the tree of life. Protoplasma 227:53-64. doi: 10.1007/s00709-005-0135-1

Zillig W (1991) Comparative biochemistry of Archaea and Bacteria. Curr Opin Genet Dev

1:544-551. doi: 10.1016/S0959-437X(05)80206-0

2.7 Figure

Fig. 1 Maximum likelihood trees of eight ARSs (SerRS, GlyRS-1, GlyRS-2, LeuRS, GluRS, TrpRS, PheRS-α, PheRS-β). These trees show a common feature that Eukaryal cytoplasmic ARS is an ingroup of Archaea. The trees were reconstructed by RAxML with optimal amino acid substitution model. Rell bootstrap support value and posterior probability are shown at the node of the root of Eukarya and the sister grouping of Eukarya. Colors of branches indicate the archaeal phylum or the domain of organisms: Red = TACK superphylum, Rose pink = Euryarchaeota, magenta = DPANN superphylum, blue = Bacteria, green = Eukarya, yellow = Eukayral organellar ARS

Fig. 2 Maximum likelihood trees of seven ARSs (ValRS, ThrRS, IleRS, AspRS, LysRS-class II, LysRS-class I, GlnRS). Monophyletic cytoplasmic ARSs in five trees (ValRS, ThrRS, IleRS, AspRS, LysRS-class II) derived from Bacteria. Numbers and colors of branches are indicated in the legend to Fig. 1

Fig. 3 Maximum likelihood trees of eight ARSs (CysRS, AsnRS, TyrRS, ProRS, HisRS, AlaRS, ArgRS, MetRS). Eight Eukarya cytoplasmic ARSs (CysRS, AsnRS, TyrRS, ProRS, HisRS, AlaRS, ArgRS, MetRS) were split into 2 or 3 groups in each of the phylogenetic trees. Numbers and colors of branches are indicated in the legend to Fig. 1

Fig. 4 The proposed universal tree of life in this study. The topologies and branch lengths in each subdomain are based on small subunit rRNA. The root branch of “Eukaryotes” was moved next to the TACK superphylum manually.

Fig. 5 The evolutionary model of Archaea to C. eukaryotes. The cytoplasm of C. eukaryotes originated from Lokiarchaeota or TACK superphylum like archaea. Then, the Lokiarchaeota or TACK superphylum like archaea accepted genes from Euryarchaeota, DPANN superphylum, and Bacteria except for α-proteobacteria via lateral gene transfer events. Lokiarchaeota or TACK superphylum like archaea also acquired mitochondria of α-proteobacteria origin and genes from α-proteobacteria via endosymbiotic gene transfer or lateral gene transfer. All of this gene flow contributes a birth of *C. eukaryotes*. A: Eukaryotic genes were derived from Lokiarchaeota or TACK superphylum like archaea. B: Eukaryotic genes were derived from Euryarchaeota. C: Eukaryotic genes derived from DPANN superphylum. D: Eukaryotic genes derived from Bacteria. The secondary candidate of ancestor of each ARS was shown in the square bracket.

Fig. 1 Maximum likelihood trees of eight ARSs (SerRS, GlyRS-1, GlyRS-2, LeuRS, GluRS, TrpRS, PheRS-α, PheRS-β).

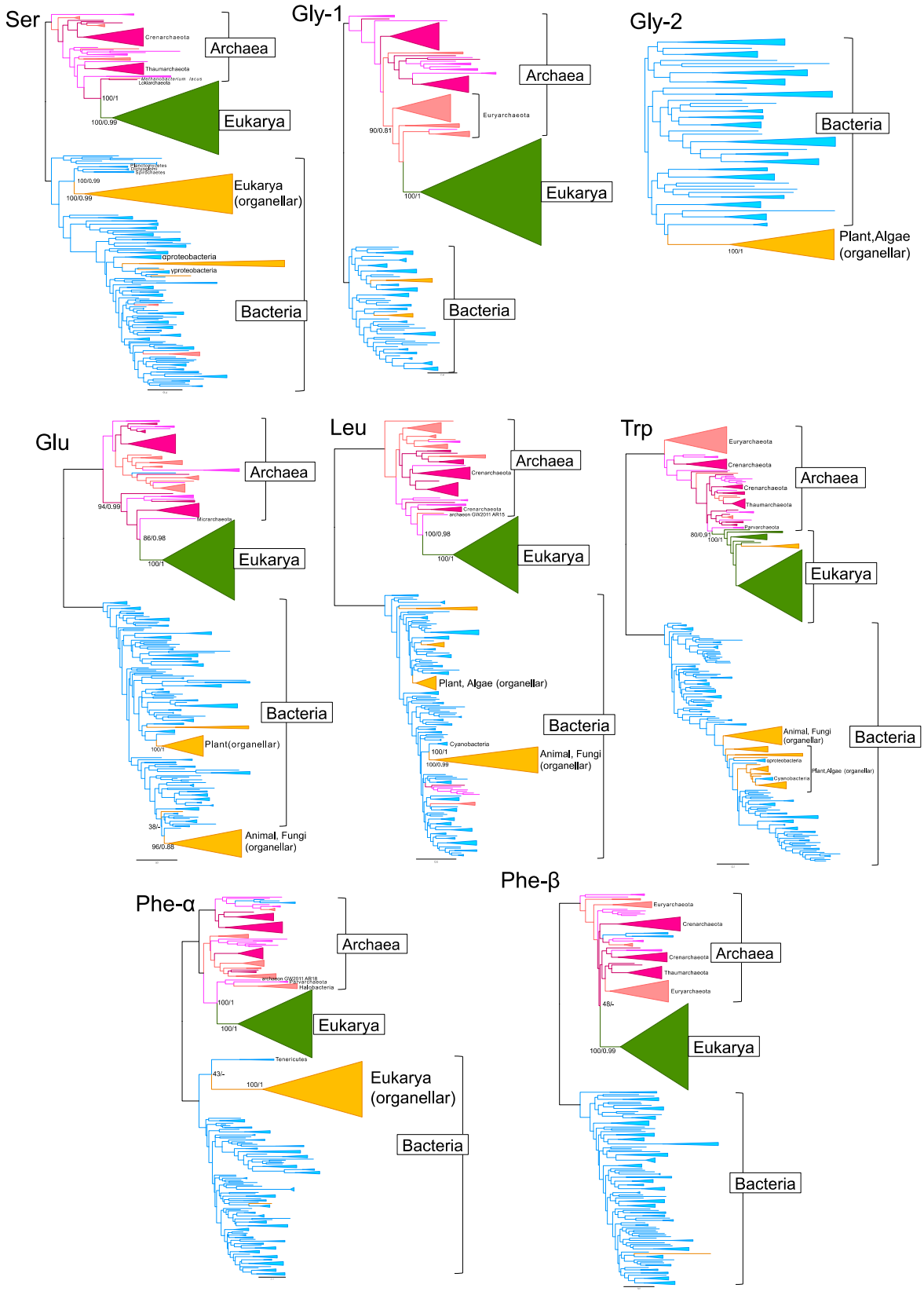


Fig. 2 Maximum likelihood trees of seven ARSs (ValRS, ThrRS, IleRS, AspRS, LysRS-class II, LysRS-class I, GlnRS)

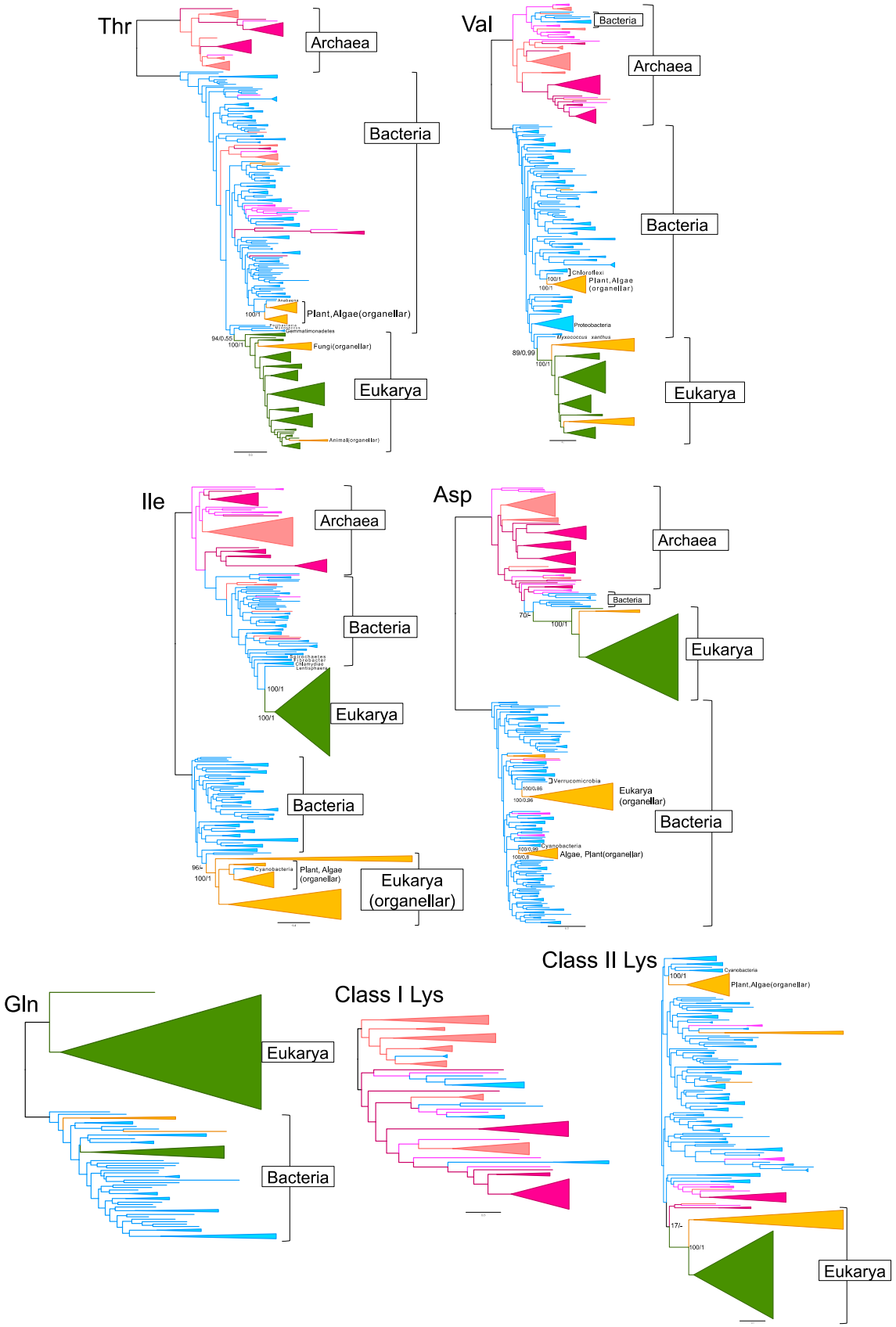


Fig. 3 Maximum likelihood trees of eight ARSs (CysRS, AsnRS, TyrRS, ProRS, HisRS, AlaRS, ArgRS, MetRS).

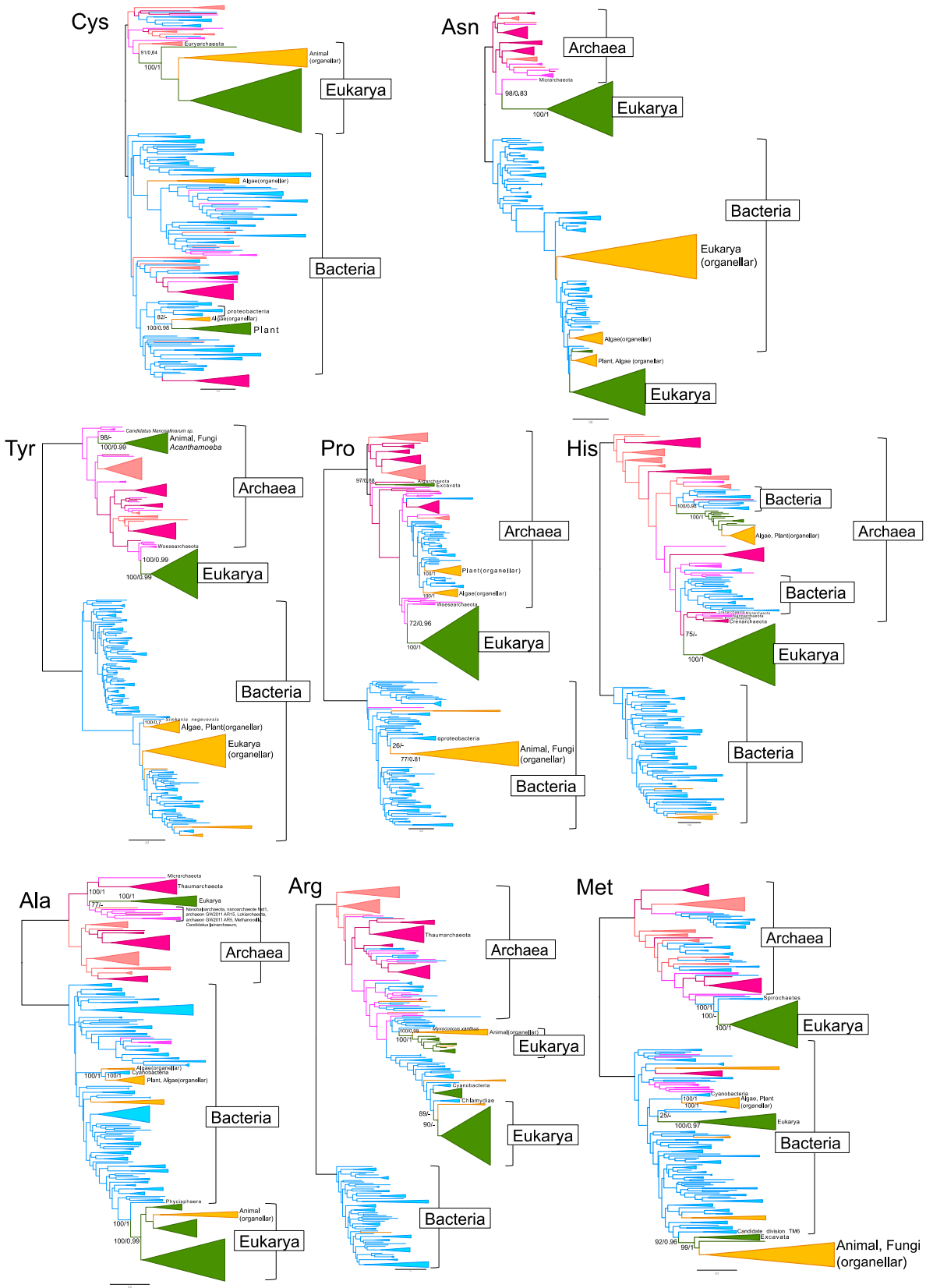


Fig. 4 The proposed universal tree of life in this study.

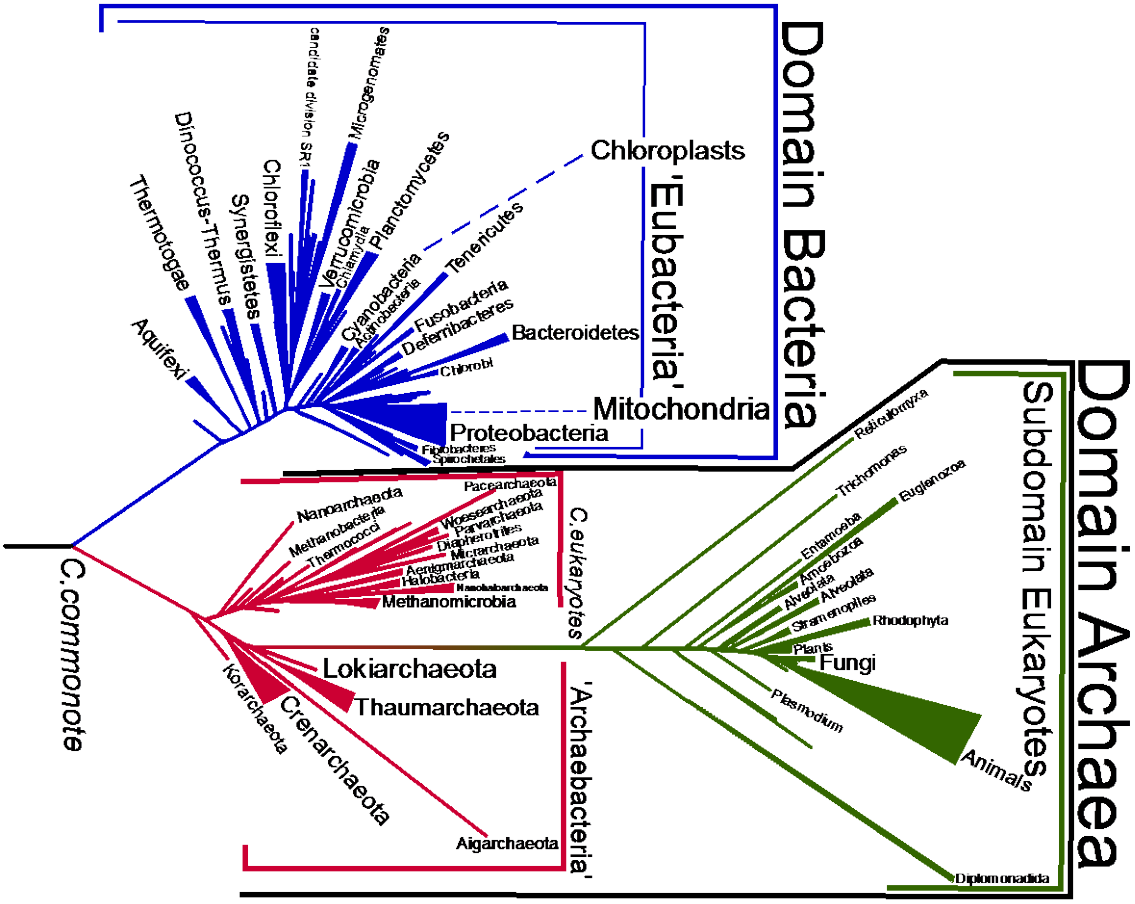
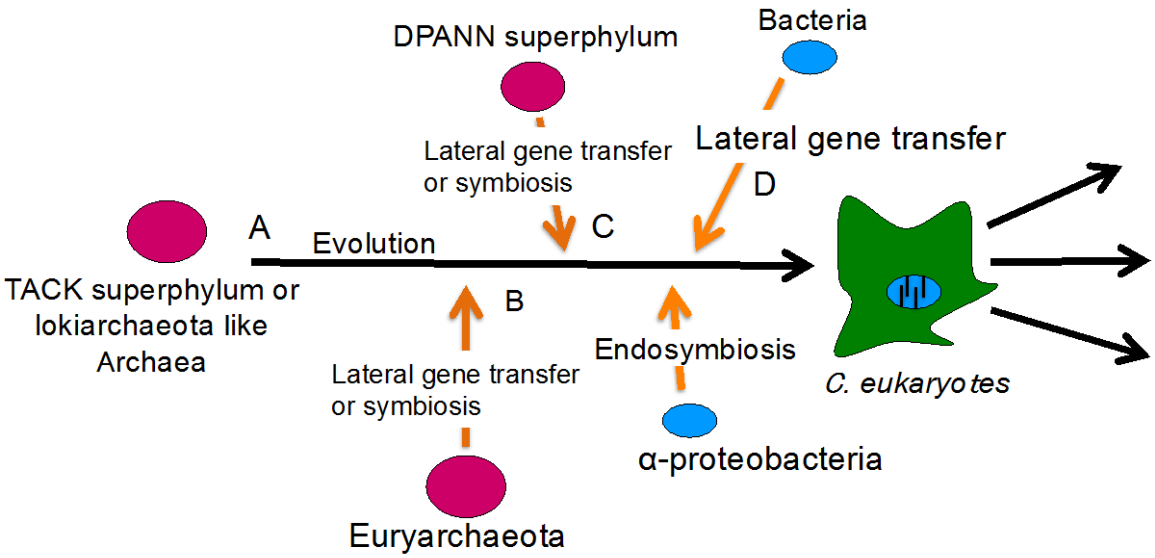


Fig. 5 Evolutionary model of Archaea to C. eukaryotes based on ARS trees.



A: SerRS, (GluRS), (LeuRS), (TrpRS), (PheRS-β), (ProRS)
B: GlyRS, CysRS, (PheRS-α), (PheRS-β)
C: GluRS, LeuRS, TrpRS, TyrRS, AsnRS, ProRS, (AlaRS), (PheRS-α)
D: ValRS, LysRS, ThrRS, IleRS, AspRS, AlaRS, MetRS, ArgRS

2.8 Table

2.8.1 Table 1. The closet Archaeal species to Eukarya in a phylogenetic tree of monophyletic eukaryal cytoplasmic ARSs.

	The closest species to Eukarya	Supporting hypothesis
SerRS	Lokiarchaeota, Methanobacterium:100/1	TACK superphylum
GlyRS-1	Euryarchaeota:90/0.81	Euryarchaeota
GluRS	Micrarchaeota (Candidatus Micrarchaeum acidiphilum ARMAN-2):86/0.98 [Thaumarchaeota]	PMW group (DPANN superphylum)
LeuRS	Woesearchaeota (Archaeon GW2011 AR15):100/0.98 [Cren., Aig., Nano., Parv., Aenigm.]	PMW group (DPANN superphylum)
TrpRS	Parvarchaeota:80/0.91 [TACK superphylum, DPANN superphylum, Thermococci]	PMW group (DPANN superphylum)
PheRS-α	Euryarchaeota (Halobacteria, Methanocella), Parvarchaeota, Woesearchaeota (Archaeon GW2011 AR18):100/1	Euryarchaeota or PMW group (DPANN superphylum)
PheRS-β	Euryarchaeota, TACK superphylum:48/-	Euryarchaeota or TACK superphylum

The numbers following the names of taxonomic groups/species are the RELL bootstrap support values in ML analysis and the posterior probability in BI analysis for eukaryal cytoplasmic ARSs being a sister group of ARSs of a certain archaeal group/species. If ARSs of a single species from an archaea phylum in which ARSs from multiple species were used in our analyses, the species name of the ARS is shown in the round bracket. The group/species names of the secondary closely related archaeal ARSs to eukaryal cytoplasmic ARSs are shown in the square bracket. TACK superphylum consists of Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota and Lokiarchaeota. The PMW group consists of Parvarchaeota, Micrarchaeota and Woesearchaeota. DPANN superphylum consists of Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota, Micrarchaeota, Pacearchaeota and Woesearchaeota.

2.8.2 Table 2. The closet Bacterial species to Eukarya in a phylogenetic tree of monophyletic eukaryal cytoplasmic ARSs.

	Eukarya evolved from	The closest species to Eukarya
ThrRS	Bacteria	Gemmatimonadetes, Deltaproteobacteria (Myxococcus), Poribacteria:94/0.55 [Chrysiogenetes]
ValRS	Bacteria	Deltaproteobacteria (Myxococcus xanthus):89/0.99 [Chrysiogenetes, proteobacteria]
IleRS	Bacteria group in Archaea	Lentisphaera:100/1 [Chlamydiae, Fibrobacter, Spirochaetes]
AspRS	Bacteria group in Archaea	Bacteria (Deinococcus-Thermus, Spirocheta, Candidatus Acetothermus, Clostridium, Microgenomates):70/-, Bacteria (Candidate division WWE3, Candidate division WS6, Peregrinibacteria):-/0.5
LysRS	Archaea or Bacteria	Archaea (Crenarchaeota, Micrarchaeota, Methanocella):17/-, Aquificae:-/0.62

The numbers following the names of taxonomic group/species are the RELL bootstrap support values in ML analysis and the posterior probability of the sister in BI analysis for eukaryal cytoplasmic ARSs being a sister group of ARSs of a certain group/species. If ARSs of a single species from a phylum/class in which ARSs from multiple species were used in our analyses, the species name of the ARS is shown in the round bracket. The group/species names of the secondary closely related ARSs to eukaryal cytoplasmic ARSs are shown in the square bracket.

2.8.3 Table 3. The closet species to Eukarya in a phylogenetic tree of polyphyletic Eukaryal cytoplasmic ARSs.

	The closest species to Eukarya			Supporting hypothesis
	1st Eukaryal group (ancestor)	2nd Eukaryal group	3rd Eukaryal group	
CysRS	Euryarchaeota (Thermococci, Methanococci):91/0.64	Proteobacteria:82/-		Euryarchaeota
AsnRS	Micrarchaeota:96/0.83	Bacteria		PMWgroup (DPANN superphylum)
ProRS	Woesearchaeota:72/0.96	Aigarchaeota:97/0.88		PMW group (DPANN superphylum)
TyrRS	Woesearchaeota:100/0.99	Candidatus Nanosalinarum:98/-, Parvarchaeota, Woesearchaeota (Archaeon GW2011 AR20):-/0.74		PMW group (DPANN superphylum)
HisRS	Crenarchaeota, Korarchaeota, Nanoarchaeota, Micrarchaeota:75/-, Peregrinibacteria:-/0.95	Fibrobacters, Verrucomicrobia, Candidate division WS6, Candidate division SR1 Saccharibacteria, Gemmatimonas, Phycisphaera, Leptospira, Lokiarchaeota:100/-, Fibrobacters:-/0.67		(TACK superphylum)
AlaRS	Phycisphaeria:100/1	Nanohaloarchaeota, Nanoarchaeote Nst1, Archaeon GW2011 AR15, Lokiarchaeota, Archaeon GW2011 AR5, Methanocella, Candidatus Iainarchaeum:77/-, Thaumarchaeota, Micrarchaeota:-/0.53		
ArgRS	Chlamydiae:89/0.72	Deltaroteobacteria (Myxococcus):100/0.99	Cyanobacteria:100/0.99	
MetRS	Spirochetes:100/-, Spirochetes, Lentisphaera:-/1	Candidate division TM6:92/0.98	Gemmatimonadetes, Latescibacteria:25/-	

The numbers following the names of taxonomic group/species are the RELL bootstrap support values in ML analysis and the posterior probability of the sister in BI analysis for eukaryal cytoplasmic ARSs being a sister group of ARSs of a certain group/species. If ARSs of a single species from a phylum/class in which ARSs from multiple species were used in our analyses, the species name of the ARS is shown in the round bracket. TACK superphylum consists of Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota and Lokiarchaeota. The PMW group consists of Parvarchaeota, Micrarchaeota and Woesearchaeota. DPANN superphylum consists of Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota, Micrarchaeota, Pacearchaeota and Woesearchaeota.

2.8.4 Table 4. Proposed higher taxonomy of life

Our Proposal		Woese et al. (1990)
Domain	Subdomain	Domain
Bacteria	Eubacteria	Bacteria
Archaea	Archaebacteria	Archaea
	Eukaryotes	Eucarya

3. Evolution of aminoacyl tRNA synthetase based on composite tree analysis

3.1 Background

Expansion of amino acid repertory in early translation system is one of the largest scientific mysteries in early evolution of life. Many hypotheses regarding evolution of genetic code have proposed on the expansion of amino acid repertory (Crick 1968; Woese 1973; Wong 1975; Eigen and Schuster 1977). Thought the order of recruitment of amino acids into the protein synthesis has been proposed (Trifonov 2004; Liu et al. 2010), no experimental or theoretical evidences has been obtained.

Aminoacyl-tRNA synthetase (ARS) is essential enzyme that attaches amino acid to cognate tRNA in translation system. The hypothesis that the expansion of ARS species has contributed the increase of amino acid repertory in early evolution of translation system has been proposed. To test this hypothesis, I reconstructed composite trees of aminoacyl-tRNA synthetase. Composite trees have been reconstructed by several groups to elucidate the position of *C. commonote* in the tree of life, previously. Iwabe et al. reconstructed composite tree of translation elongation factor and propose that the root of universal tree is placed between Bacteria and the common ancestor of Archaea and Eukarya (1989). Brown et al. constructed composite trees of IleRS, ValRS and LeuRS (Brown and Doolittle 1995), which suggested the root of all extant organisms placed between Bacteria and the common ancestor of Archaea and Eukarya. Kollman and Doolittle (2000) reconstructed composite trees of TyrRS/TrpRS, SerRS/ThrRS. Both trees also suggested the root of all extant organisms placed between Bacteria and the common ancestor of Archaea and Eukarya. Zhaxybayeva et al. (2005) reviewed the paralogous rooting using composite trees, which suggested that the majority of datasets supported that the root position of *C. commonote* between Bacteria and Archaea. However, taxonomic abundance is not sufficient in these analyses. To clarifying the detailed evolutionary history, specifically the root of all extant organisms, phylogenetic analysis using abundant taxonomical species is needed. In this chapter, I focused that the root of each ARS in the composite tree of each subclass and examined the order of expansion of amino acid repertory.

3.2 Materials and Methods

3.2.1 Sequence Data of ARS

We selected two or three typical species from each order to reduce taxonomic bias. All protein sequences of 118 selected organisms (Archaea: 23, Bacteria: 57, Eukarya: 38) were collected from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). We constructed a KF database (M. Kanetake, R. Furukawa, S. Yokobori, and A. Yamagishi, unpublished) in Geneious ver. 7.1.9 (http://www.geneious.com, Kearse et al. 2012) that consisted of all protein sequences of 118 organisms. The KF database was first constructed on 14 October 2010. Protein sequences of 23 ARSs were searched with BlastP (Altschul et al. 1997) from the KF database. Accession numbers of all collected data is shown in Supplemental table S4.

3.2.2 Sequence Alignment

Collected amino acid sequences were classifies into each subclass of ARS (class Ia [MetRS, ValRS, LeuRS, IleRS, CysRS and ArgRS], class Ib [GluRS, GlnRS and LysRS-class I], class Ic [TyrRS and TrpRS], class IIa [SerRS, ThrRS, AlaRS, GlyRS-α_2, ProRS and HisRS], class IIb [AspRS, AsnRS and LysRS-class II], and class IIc [PheRS, GlyRS-$\alpha_2\beta_2$]). Recent study of RMSD cluster dendrogram of ARS proposed that PheRS, SepRS and PylRS are classified into class IIc and that AlaRS and GlyRS-2 are classified into class IId (Valencia-Sánchez et al. 2016). Referring to new classification, AlaRS was removed from class IIa ARS and classified into class IId ARS with GlyRS-2. Amino acid sequences of each subclass ARS were aligned using MAFFT 7.017 (Katoh et al. 2013) and edited manually. As a result, seven composite alignments of ARS were constructed. The editing domain in bacterial LeuRS is located after the ZN-1 domain, whereas the editing domain is located before the ZN-1 domain in archaeal/eukaryal LeuRS, IleRS and ValRS (Cusack et al. 2000). The editing domain in bacterial LeuRS was transferred in front of the ZN-1 domain during the manual alignment. The well-aligned regions of each alignment were selected from the final alignment using TrimAl 1.4 (Capella-Gutierrez et al. 2009). TrimAl was used with automated1 mode and the columns containing gap were excluded with nogaps mode. The numbers of sites of the final alignment of 23 ARSs are shown in supplemental table S5.

3.2.3 Phylogenetic analysis

The optimal amino acid substitution model for each composite alignment was selected using the model selection program PROTTEST 3.4 (Darriba et al. 2011) and is shown

in supplemental table S5. Composite trees of each subclass of ARS were reconstructed by Maximum Likelihood (ML) and Bayesian Inference (BI) analyses. ML analyses were done with the program RAxML 8.1 (Stamatakis 2014) with optimal amino acid substitution model for each ARS. Bootstrap analysis was done by analyzing 100 resampled data sets. Posterior-probability consensus trees in BI analysis were constructed using PhyloBayes 3.3f (Lartillot et al. 2009) by running two chains until the max discrepancy dropped lower than 0.3 under the CAT Poisson + Γ(4) model. The consensus tree was output using the readpb program. The trees used to readpb analysis were sampled every 10 generations in each analysis. The number of cut-off trees and the reached generation of chains in each analysis are shown in supplemental table S5.

3.3 Result and Discussion

3.3.1 Class Ia aminoacyl tRNA synthetase

Composite trees of class Ia ARS (IleRS, LeuRS, ValRS, MetRS, CysRS, ArgRS) were reconstructed by ML and BI methods (Fig.6, 7 and Supplemental Figure S4). The root of the tree was placed between ArgRS and other ARSs following the result of Nagel and Doolittle (1995). CysRS diverged earliest in both ML and BI trees. Monophyletic group of MetRS diverged second earliest in BI tree. However in ML tree, bacterial MetRS and and archaeal/bacterial/eukaryal MetRS diverged second and third earliest, respectively. Archaeal/eukaryal LeuRS diverged and bacterial LeuRS diverged one by one in both trees. Archaeal ValRS diverged next in ML tree. However in BI tree, monophyletic group of ValRS and IleRS diverged next. The monophyletic group of bacterial ValRS and IleRS diverged last was supported with 48% bootstrap value (bp) in ML tree. These trees raised the questions regarding the respective monophyly of MetRS, LeuRS and ValRS.

The root position of ArgRS was in Bacteria in both trees. The root position of CysRS was in archaeal CysRS in both trees. The root position of MetRS was between Bacteria and Archaea/Eukarya groups in ML tree. The root position of ValRS in BI tree was in archaeal group. While, the root position of IleRS was between Bacteria and Archaea/Eukarya in ML tree, which is in archaeal group in BI tree. These results are consistent with the composite tree of IleRS, LeuRS, ValRS and MetRS in my master thesis (2012). However, the root position of ValRS and IleRS in BI tree was inconsitent with the current ML tree and previous composite tree (supplemental Figure S5, Furukawa master thesis 2012).

The difference in topology of MetRS, ValRS and IleRS in six ARSs composite tree in this thesis (Figs. 6 and 7) and the four ARSs composite tree in my master theis (supplemental Figure S5, Furukawa master thesis 2012) was examined. The root position of MetRS and IleRS was between Bacteria and Archaea/Eukarya group in previous composite tree, which agree the MetRS in BI tree and IleRS in ML tree. LeuRS showed paraphyletic topology similar to that in six ARSs tree. ValRS showed paraphyletic topology similar to that in ML tree of six ARSs. In previous studies (Brown and Doolittle 1995, Fournier et al. 2011) ValRS and LeuRS showed monophyly. To test the monophyly of ValRS and LeuRS, I performed hypothesitical test. Hypothesitical test supported the paraphyletic topology of ValRS and paraphyletic topology of LeuRS the best. However, monophyly of ValRS and the monophyly of LeuRS was not rejected completely.

When the expantion of amino acid repertory depend on the evolution of ARS, this

composite tree suggest the order of expantion of amino acid repertory. Composite tree of class Ia shows that arginine appearred earliest and cysteine, methionine, leucine, valine and isoleucine appeared in the order. Fournier et al. estimated composite tree of LeuRS, ValRS and IleRS and paralog ancestral sequence of ValRS and IleRS (2011). Frequence of aliphatic amino acid (leucine, valine, isoleucine) in paralog ancestral sequence of ValRS and IleRS showed similar overall counts, which suggested that leucine, valine and isoleucine were used in aminoacylation before appearance of ValRS and IleRS (Fournier et al. 2011). They also suggested that the genetic code did not co-evolve with the ARS in this divergent case (Fournier et al. 2011). To clarify the expantion of amino acid repertory, biochemical experiment is on going by the other member of our group.

3.3.2 Class Ib aminoacyl tRNA synthetase

Composite trees of class Ib ARS (LysRS, GluRS and GlnRS) were reconstructed by ML and BI methods (Figs. 8, 9 and Supplemental Figure S4). The root of the tree was placed between LysRS and GluRS based on the previous composite tree analyses (Ribas de Pouplana et al. 1998; Nureki et al. 2010). Monophyly of each ARS (LysRS and GluRS/GlnRS) is supported with 100% bp in ML tree and 1.00 pp in BI tree, which is consistent with previous studies (Ribas de Pouplana et al. 1998; Nureki et al. 2010). The root position of LysRS was in archaeal group in both analyses. The result is related to the situation that archaeal and bacterial groups have LysRS in different class, class Ib and class IIb, respectively. The root position of GluRS was between Bacteria and Archaea/Eukarya group in both analyses. Thus, the root position of GluRS supports the position of *C. commonote* between Bacteria and Archaea/Eukarya.

GlnRS was ingroup of archaeal GluRS and was a sister group of Eukaryal GluRS. Evolutionary history of GlnRS was already suggested that GlnRS evolved by a gene duplication of the eukaryal GluRS and lateral gene transfer from early eukarya to some bacterial group (Lamoure et al. 1994; Ribas de Pouplana et al. 1998; Siatecka 1998; Brown and Doolittle 1999; Woese et al. 2000; Nureki et al. 2010). Most Bacteria and Archaea have no GlnRS and use Glu-tRNA^{Gln} amidotransferase (Glu-AdT) that allows the formation of correctly charged Gln-tRNA^{Gln} (Cathopoulis et al. 2007). This Glu-AdT is distributed in most Bacteria, Archaea and Eukarya species, which suggest that GlnRS was late invention evolved from GluRS (Sheppard and Söll 2008). My result is consistent with these reports.

3.3.3 Class Ic aminoacyl tRNA synthetase

Composite trees of class Ic ARS (TyrRS and TrpRS) were reconstructed by ML and BI analyses (Fig. 10, 11 and Supplemental Figure S4). The root of the tree was placed between TyrRS and TrpRS. Monophyly of each ARS was supported with 97% bp in ML tree and 0.99 pp in BI tree, which is consistent with previous studies (Brown et al. 1997; Fournier and Alm 2015). The root position of TyrRS was between Bacteria and Archaea/Eukarya groups in both ML and BI analyses, which is consistent with previous studies (Brown et al. 1997; Fournier and Alm 2015). In ML tree, the root of TrpRS was between Bacteria and Archaea/Eukarya group, which is also consistent with previous studies (Brown et al. 1997; Fournier and Alm 2015). However, in BI tree, the root of TrpRS was within Bacteria. Archaeal group also diverged from Bacteria supported with 0.50 pp. Since the resolution of deep branch of TrpRS in BI analysis is very low, the root position in ML tree is more probable. Accordingly, TyrRS-TrpRS composite tree supports the position of *C. commonote* between Bacteria and Archaea/Eukarya groups. Internal phylogenetic relationship of each ARS was consistent with our single gene phylogenetic tree (Figs. 1 and 3).

3.3.4 Class IIa aminoacyl tRNA synthetase

Composite trees of class IIa ARS (HisRS, GlyRS-1 ThrRS, ProRS and SerRS) were reconstructed by ML and BI methods (Fig. 12, 13 and Supplemental Figure S4). The root of the tree was placed between HisRS and other ARS species based on the composite tree of class IIa ARS and class IIb ARS (supplemental Figure S6). In ML tree, GlyRS-1 diverged earliest and ThrRS diverged second earliest. However, ThrRS diverged earliest and GlyRS-1 diverged second earliest in BI tree. ML tree showed similar topology with composite tree of class IIab ARS (Supplemental Figure S6). Accordingly, the topology of ML tree where GlyRS-1 diverged earliest has higher probability. Monophyletic relationship of ProRS and SerRS was supported 45% bp and 0.99 pp. This monophyletic relationship was also supported in composite tree of class IIa ARS and class IIb ARS (Supplemental Figure S6).

Monophyly of HisRS was supported with 100% bp in ML tree and 1.00 pp in BI tree. Monophyly of GlyRS-1 is also supported with 95% bp and 1.00 pp. Monophyly of ThrRS was supported with 72% bp in ML tree and 0.88 pp. Monophyly of ProRS was supported with 61% in ML tree, but in BI tree ProRS showed paraphyletic topology where archaeal ProRS diverged earlier than bacterial ProRS. Though, monophyly of ProRS was obtained in the class IIab composite ML tree (Supplemental Figure S6a), the monophyly of ProRS was ambiguous in class IIab composite BI tree (Supplemental Figure S6b). To clarify whether the monophyly of ProRS, phylogenetic analyses using ProRS with an outgroup ARS is required. Monophyly of SerRS was supported with 56% bp and 0.99 pp.

3.3.5 Class IIb aminoacyl tRNA synthetase

Composite trees of class IIb ARS (LysRS, AspRS and AsnRS) were reconstructed by ML and BI methods (Fig.14, 15 and Supplemental Figure S4). The root of the tree was placed between LysRS and AspRS based on previous composite tree (Nair et al. 2016). Monophyly of each ARS (LysRS and AspRS/AsnRS) was supported with 100% bp in ML tree and 1.00 pp in BI tree, which is consistent with previous studies (Nair et al. 2016). The root position of LysRS was in bacterial group in both analyses, because the archaeal group has class Ib LysRS. The root position of AspRS was between Bacteria and Archaea/Eukarya groups in ML analyses. On the other hand, the root position of AspRS in BI analaysis was in eukaryal organellar group, and archaeal group diverged from bacterial group. Since the resolution of deep branch of AspRS in BI analysis is very low, the root position of ML tree is more reliable. Accordingly, class IIb this composite tree supported the position of *C. commonote* between Bacteria and Archaea/Eukarya groups.

AsnRS was ingroup of archaeal AspRS in both analyses, which support the evolutionary history that AsnRS was derived from archaeal AspRS. This result support that AsnRS originated from AspRS, with a little difference from previous phylogenetic studies which suggest that AsnRS is a sister group of archaeal AspRS or archaeal/eukaryal AspRS (Woese et al. 2000; Charron et al. 2003; Roy et al. 2003; Charrière et al. 2009; Nair et al. 2016). My result is more accurate than previous studies because I used the adequate amino acid substitution model with larger number of sequence entries for phylogenetic analysis.

Gene duplication of archaeal AspRS occurred in early archaeal lineage followed by the formation of AsnRS. Archaeal AsnRS gene was transferred to bacteria. Eukaryal AsnRS was derived from archaeal AsnRS through vertical evolution independent from gene transfer from Archaea to Bacteria.

Bacteria or Archaea without AsnRS use Asp-tRNA^{Asn} amidotransferase (Asp-AdT) that allows the formation of correctly charged Asn-tRNA^{Asn} (Cathopoulis et al. 2007). This Asp-AdT is distributed in most Bacteria, Archaea and Eukarya species, which suggest that AsnRS were late inventions, evolving from AspRS (Sheppard and Söll 2008). *C. commonote* must have used Asp-AdT.

3.3.6 Class IIc aminoacyl tRNA synthetase

PheRS is a heterotetramer consisting of two short α subunits and two long β subunits. PheRS-α and PheRS-β show very low sequence similarities with typical 3 motifs of class II ARS. Pyramidal classification of PheRS-α and PheRS-β suggested that the PheRS-β probably arose from the duplication of an ancestral catalytic domain of class II ARS followed by subsequent insertions and deletions of polypeptides (Diaz-Lazcoz et al. 1998).

Composite trees of class IIc ARS (PheRS-α/PheRS-β) were reconstructed by ML and BI methods (Fig. 16, 17 and Supplemental Figure S4). If the root of the composite tree of class IIc ARS was placed between PheRS-α and PheRS-β, monophyly of each ARS (PheRS-α, PheRS-β) is supported with 100% bp in ML tree and 1.00 pp in BI tree, which is consistent with previous studies (Diaz-Lazcoz et al. 1998, Lin and Huang 2003). The root position of PheRS-β is in bacterial group in both ML and BI analyses with low resolution. The root position of PheRS-α is between Bacteria and Archaea/Eukarya group in both ML and BI analyses. Thus, PheRS-α supports the position of *C. commonote* between Bacteria and Archaea/Eukarya group, though the resolution of PheRS-β is too low to judge the topology.

However, this composite tree was reconstructed from short conserved aligned sites (Supplemental Table S5). Thus, improvement of alignment method and searching for similar protein with PheRS-α or PheRS-β as the out group for composite tree is needed to understand the evolution of each subunit of PheRS.

3.3.7 Class IId aminoacyl tRNA synthetase

The composite tree of class IId (AlaRS/GlyRS-2) was reconstructed by ML and BI methods (Fig.18, 19 and Supplemental Figure S4). If the root of the composite tree of class IId ARS was placed between AlaRS and GlyRS-2, monophyly of each ARS (AlaRS and GlyRS-2) was supported with 100% bp in ML tree and 1.00 pp in BI tree. The root position of GlyRS-2 was in bacterial group in both analyses, because GlyRS-2 can be found only in bacterial group. The root position of AlaRS was between Bacteria and Archaea/Eukarya group in ML tree. However in BI tree, the root position of AlaRS was in archaeal group.

Based on the structure analyses Valencia-Sánchez et al. proposed the hypothesis that GlyRS-2 was derived from AlaRS and GlyRS-1 originated from unrelated ancestor to GlyRS-2 (2016). Two GlyRSs have totally different origins based on sequence and structure analyses (Valencia-Sánchez et al. 2016). My ML tree suggests that AlaRS and GlyRS-2 diverged from common ancestor. If the root of the tree was placed between bacterial AlaRS and archaeal AlaRS in BI tree, GlyRS-2 is diverged from archaeal AlaRS, which support the late origin of GlyRS-2. To clarify the true evolutional story, I will perform phylogenetic analyses using these 2 ARSs (AlaRS and GlyRS-2) with other class II ARS.

3.4 Conclusion

I have reconstructed 14 composite trees. The root position is summarized in Table 5. Some of the root had low resolution. Some ARSs appeared later in the history after the divergence of *C. commonote*. Two types of ARS can be found for one cognate amino acid, for GlyRS and LysRS. The root positions of these ARSs cannot be used for the determination of the position of *C. commonote*. The reliable root position in my seven composite trees showed the root position of the *C. commonote* is between Bacteria and Archaea in 14 cases (Table 5).

The order of incorporation of amino acid species in protein synthesis has been proposed based on tendency of amino acid abundance in the history after *C. commonote* (Liu et al. 2010). Though it is possible to find the order of branching of each ARS species in these composite trees, it is not directly related to the amino acid species used at the branching point: Both amino acid species used after the divergence may be used at the branching point. However, it may be possible to check the amino acid specificity of the ancestral ARS corresponding the branching point of the two ARS species. The resurrection and analysis of the ancestral ARS is on going in other members in my lab.

Composite trees of each class have been reconstructed using less species (Nagel and Doolittle 1991; 1995). Structure dendrogram of each class was reconstructed (Donoghue et al. 2003). These analyses provided important information that we can trace back the ancestor of class I ARS and the ancestor of class II ARS. Though increasing number of ARS data are available, the detail composite trees of ARS of each class have not been reported. Although Andam and Gogarten have reported composite tree of class II ARS, they have used less number of species (2011). Accordingly, detailed composite tree with more taxonomical entries is needed to clarify ARS evolution. Aravind et al. have suggested that the catalytic domain of class I ARS is conserved as Rossmann-like topology in another proteins and the ancestor of class I ARS is diverged from primitive protein in RNA world (2003). Tracing back to the ancestor of ARS of each class will lead us to the primitive translation system and the era of protein emerged in RNA world.

3.5 Reference

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402. doi: 10.1093/nar/25.17.3389

Andam, CP, Gogarten JP (2011) Biased gene transfer and its implications for the concept of lineage. Biology direct 6:1.

Aravind L, Anantharaman V, Koonin EV (2002) Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA world. Proteins: Structure, Function, and Bioinformatics 48:1-14.

Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci USA 92:2441-2445. doi: 10.1073/pnas.92.7.2441

Brown, JR, Robb FT, Weiss R, Doolittle WF (1997) Evidence for the early divergence of tryptophanyl-and tyrosyl-tRNA synthetases. *Journal of molecular evolution*, *45*:9-16.

Brown JR, Doolittle WF (1999) Gene descent, duplication, and horizontal transfer in the evolution of glutamyl-and glutaminyl-tRNA synthetases. J Mol Evol 49:485-495. doi: 10.1007/PL00006571

Cathopoulis T, Chuawong P, Hendrickson TL (2007) Novel tRNA aminoacylation mechanisms. Molecular BioSystems, 3:408-418.

Charriere F, O'Donoghue P, Helgadóttir S et al (2009) Dual targeting of a tRNAAsp requires two different aspartyl-tRNA synthetases in Trypanosoma brucei. Journal of Biological Chemistry 284:16210-16217.

Charron C, Roy H, Blaise M, Giegé R, Kern D (2003) Non-discriminating and discriminating aspartyl-tRNA synthetases differ in the anticodon-binding domain. The EMBO journal 22:1632-1643.

Cusack S, Yaremchuk A, Tukalo M (2000) The 2 Å crystal structure of leucyl - tRNA synthetase and its complex with a leucyl - adenylate analogue. EMBO J 19:2351-2361. doi: 10.1093/emboj/19.10.2351

Crick FH (1968) The origin of the genetic code. Journal of molecular biology 38:367-379.

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164-1165. doi: 10.1093/bioinformatics/btr088

De Pouplana LR, Turner RJ, Steer BA, Schimmel P (1998) Genetic code origins: tRNAs older than their synthetases? Proceedings of the National Academy of Sciences 95:11295-11300.

Diaz-Lazcoz Y, Aude JC, Nitschke P, Chiapello H, Landes-Devauchelle C, Risler JL (1998) Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. Molecular biology and evolution 15:1548-1561.

Eigen M, Schuster P (1977) A principle of natural self-organization. Naturwissenschaften, 64:541-565.

Fournier GP, Andam CP, Alm EJ, Gogarten JP (2011) Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. Origins of Life and Evolution of Biospheres, 41:621-632.

Fournier GP, Alm EJ (2015) Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. *Journal of molecular evolution, 80*:171-185.

Iwabe N, Kuma KI, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355-9359.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780. doi: 10.1093/molbev/mst010

Kearse M, Moir R, Wilson A et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649. doi: 10.1093/bioinformatics/bts199

Kollman JM, Doolittle RF (2000) Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. Journal of Molecular Evolution 51:173-181.

Korencic D, Polycarpo C, Weygand-Durasevic I, Söll D (2004) Differential modes of transfer RNASer recognition in Methanosarcina barkeri. Journal of Biological Chemistry 279:48780-48786.

Lamour V, Quevillon S, Diriong S, N'guyen VC, Lipinski M, Mirande MARC (1994) Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of

horizontal gene transfer. Proc Natl Acad Sci U S A 91:8670-8674. doi: 10.1073/pnas.91.18.8670

Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288. doi: 10.1093/bioinformatics/btp368

Lin J, Huang JF (2003) Evolution of phenylalanyl-tRNA synthetase by domain losing. ACTA BIOCHIMICA ET BIOPHYSICA SINICA-CHINESE EDITION-, 35:1061-1065.

Liu CC, Schultz PG (2010) Adding new chemistries to the genetic code. Annual review of biochemistry 79:413-444.

Nagel GM, Doolittle RF (1991) Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proceedings of the National Academy of Sciences 88:8121-8125.

Nair N, Raff H, Islam MT, Feen M, Garofalo DM, Sheppard K (2016) The Bacillus subtilis and Bacillus halodurans Aspartyl-tRNA Synthetases Retain Recognition of tRNA Asn. Journal of molecular biology, 428:618-630.

Nureki O, O'Donoghue P, Watanabe N, Ohmori A, Oshikane H, Araiso Y, ... & Ishitani R (2010). Structure of an archaeal non-discriminating glutamyl-tRNA synthetase: a missing link in the evolution of Gln-tRNAGln formation. *Nucleic acids research*, *38*:7286-7297.

O'Donoghue P, Luthey-Schulten Z (2003) On the evolution of structure in aminoacyl-tRNA synthetases. Microbiology and Molecular Biology Reviews 67:550-573.

Roy H, Becker HD, Reinbolt J, Kern D (2003) When contemporary aminoacyl-tRNA synthetases invent their cognate amino acid metabolism. Proceedings of the National Academy of Sciences 100:9837-9842.

Sheppard K, Söll D (2008) On the evolution of the tRNA-dependent amidotransferases, GatCAB and GatDE. Journal of molecular biology 377:831-844.

Siatecka M, Rozek M, Barciszewski J, Mirande M (1998) Modular evolution of the Glx - tRNA synthetase family. Eur J Biochem 256:80-87. doi: 10.1046/j.1432-1327.1998.2560080.x

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313. doi: 10.1093/bioinformatics/btu033

Trifonov EN (2004) The triplet code from first principles. Journal of Biomolecular structure and dynamics 22:1-11.

Valencia-Sánchez MI, Rodríguez-Hernández A, Ferreira R et al (2016). Structural Insights into the Polyphyletic Origins of Glycyl tRNA Synthetases. Journal of Biological Chemistry,

jbc-M116.

Zhaxybayeva O, Lapierre P, Gogarten JP (2005) Ancient gene duplications and the root (s) of the tree of life. Protoplasma 227:53-64. doi: 10.1007/s00709-005-0135-1

Woese CR (1973) Evolution of the genetic code. Naturwissenschaften, 60:447-459.

Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64: 202-236. doi: 10.1128/MMBR.64.1.202-236.2000

Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 9:689-710. doi: 10.1101/gr.9.8.689

Wong JTF (1975) A co-evolution theory of the genetic code. Proceedings of the National Academy of Sciences of the United States of America 72:1909.

3.6 Figure

Fig. 6 Maximum likelihood composite tree of class Ia ARSs (ArgRS, CysRS, MetRS, LeuRS, ValRS, IleRS). This tree is rooted between ArgRS and other ARSs. The tree was reconstructed by RAxML with optimal amino acid substitution model. Bootstrap support value is shown at the node of the root of each ARS. Colors of branches indicate the archaeal phylum or the domain of organisms: Red = Archaea, blue = Bacteria, green = Eukarya, yellow = Eukayral organellar ARS

Fig. 7 Bayesian composite tree of class Ia ARSs (ArgRS, CysRS, MetRS, LeuRS, ValRS, IleRS). This tree is rooted between ArgRS and other ARSs. The tree was reconstructed using PhyloBayes. Posterior probability and posterior probability are shown at all nodes. Colors of names indicate the domain of organisms: Red = Archaea, blue = Bacteria, green = Eukarya, yellow = Eukayral organellar ARS.

Fig. 8 Maximum likelihood composite tree of class Ib ARSs (GluRS, LysRS-class I, GlnRS). This tree is rooted between LysRS-class I and GluRS. Numbers and colors of branches are indicated in the legend to Fig. 6

Fig. 9 Bayesian composite tree of class Ib ARSs (GluRS, LysRS-class I, GlnRS). This tree is rooted between LysRS-class I and GluRS. Numbers and colors of branches are indicated in the legend to Fig. 7

Fig. 10 Maximum likelihood composite tree of class Ic ARSs (TrpRS, TyrRS). This tree is rooted between TyrRS and TrpRS. Numbers and colors of branches are indicated in the legend to Fig. 6

Fig. 11 Bayesian composite tree of class Ic ARSs (TrpRS, TyrRS). This tree is rooted between TyrRS and TrpRS. Numbers and colors of branches are indicated in the legend to Fig. 7

Fig. 12 Maximum likelihood composite tree of class IIa ARSs (HisRS, GlyRS-1, ThrRS, ProRS, SerRS). This tree is rooted between HisRS and other ARSs. Numbers and colors of branches are indicated in the legend to Fig. 6

Fig. 13 Bayesian composite tree of class IIa ARSs (HisRS, GlyRS-1, ThrRS, ProRS, SerRS). This tree is rooted between HisRS and other ARSs. Numbers and colors of branches are indicated in the legend to Fig. 7

Fig. 14 Maximum likelihood composite tree of class IIb ARSs (AspRS, LysRS-class I, AsnRS). This tree is rooted between LysRS-class II and AspRS. Numbers and colors of branches are indicated in the legend to Fig. 6

Fig. 15 Bayesian composite tree of class IIb ARSs (AspRS, LysRS-class I, AsnRS). This tree is rooted between LysRS-class II and AspRS. Numbers and colors of branches are indicated in the legend to Fig. 7

Fig. 16 Maximum likelihood composite tree of class IIc ARSs (PheRS-α, PheRS-β). This tree is rooted between PheRS-α and PheRS-β. Numbers and colors of branches are indicated in the legend to Fig. 6

Fig. 17 Bayesian composite tree of class IIc ARSs (PheRS-α, PheRS-β). This tree is rooted between PheRS-α and PheRS-β. Numbers and colors of branches are indicated in the legend to Fig. 7

Fig. 18 Maximum likelihood composite tree of class IId ARSs (AlaRS, GlyRS-2). This tree is rooted between AlaRS and GlyRS-2. Numbers and colors of branches are indicated in the legend to Fig. 6

Fig. 19 Bayesian composite tree of class IId ARSs (AlaRS, GlyRS-2). This tree is rooted between AlaRS and GlyRS-2. Numbers and colors of branches are indicated in the legend to Fig. 7

Fig. 6 Maximum likelihood composite tree of class Ia ARSs

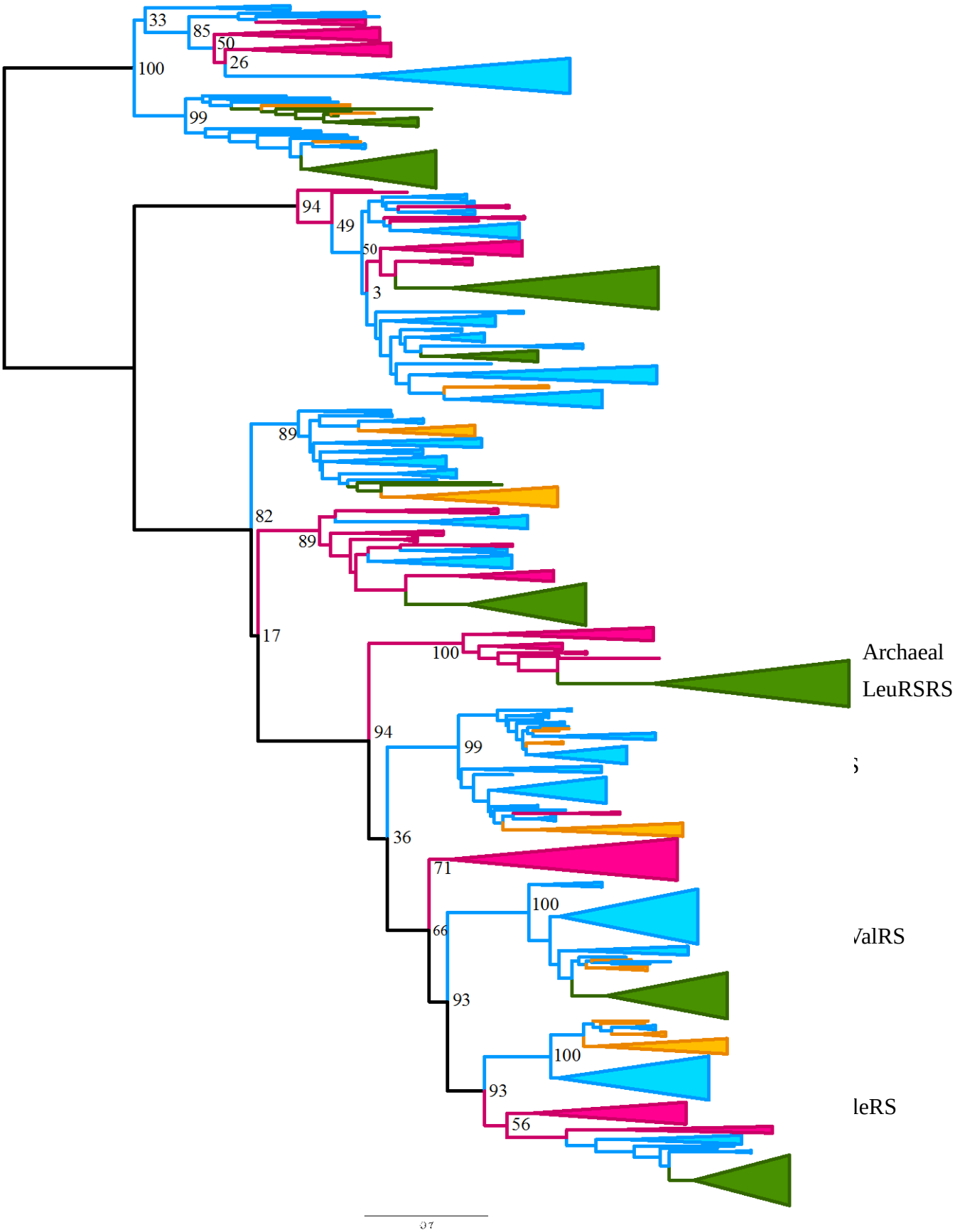


Fig. 7 Bayesian composite tree of class Ia ARSs

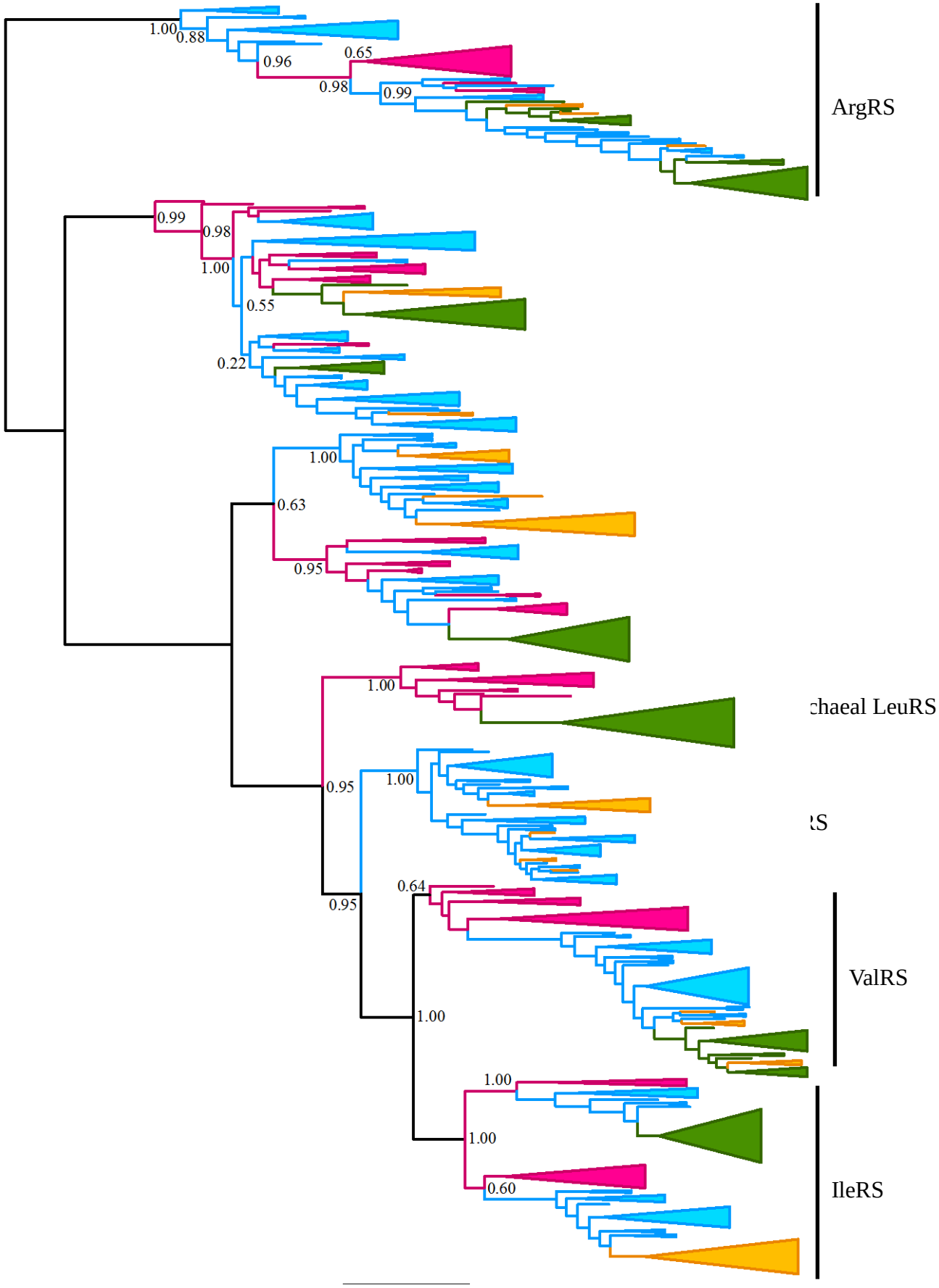


Fig. 8 Maximum likelihood composite tree of class Ib ARSs

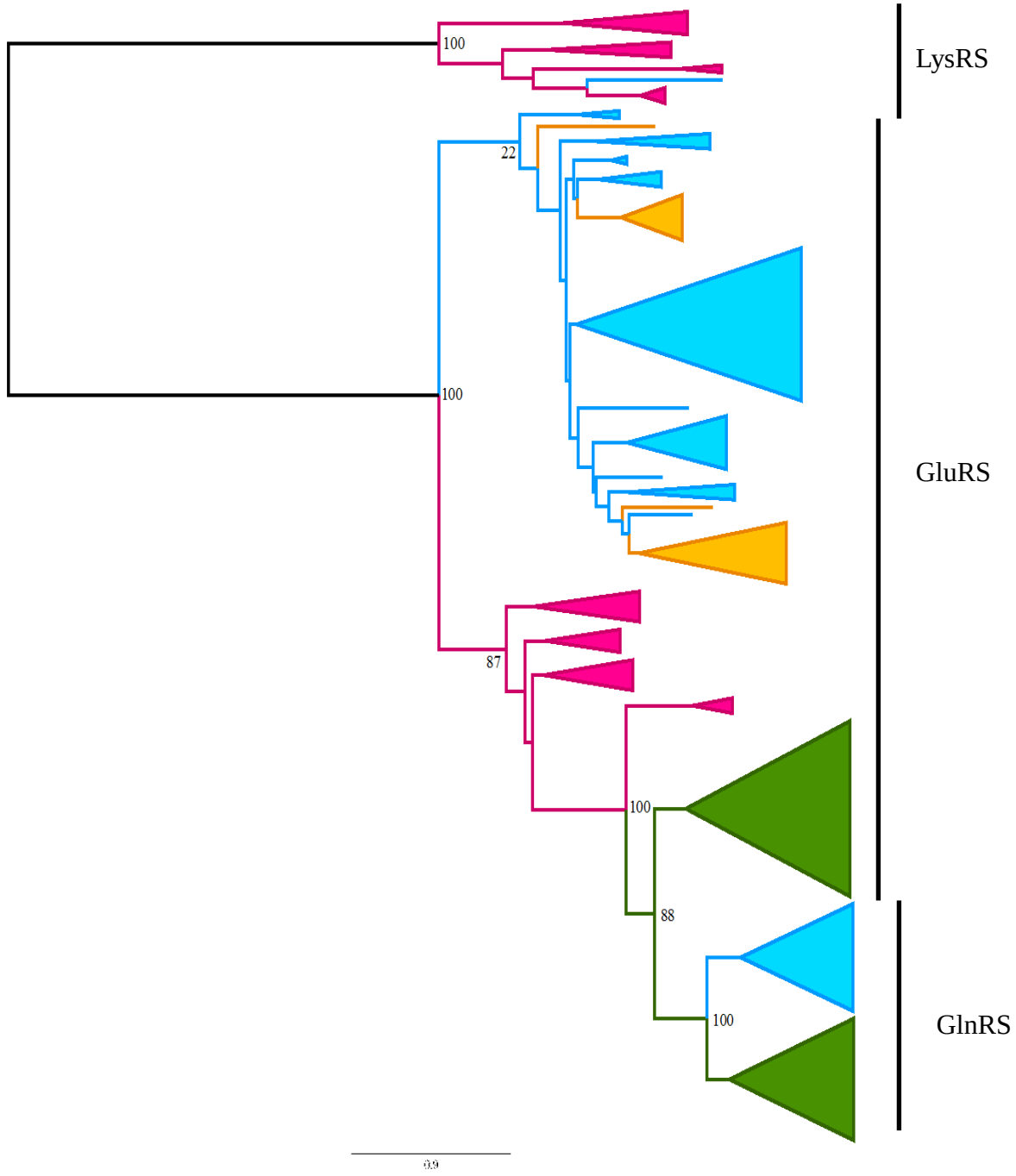


Fig. 9 Bayesian composite tree of class Ib ARSs

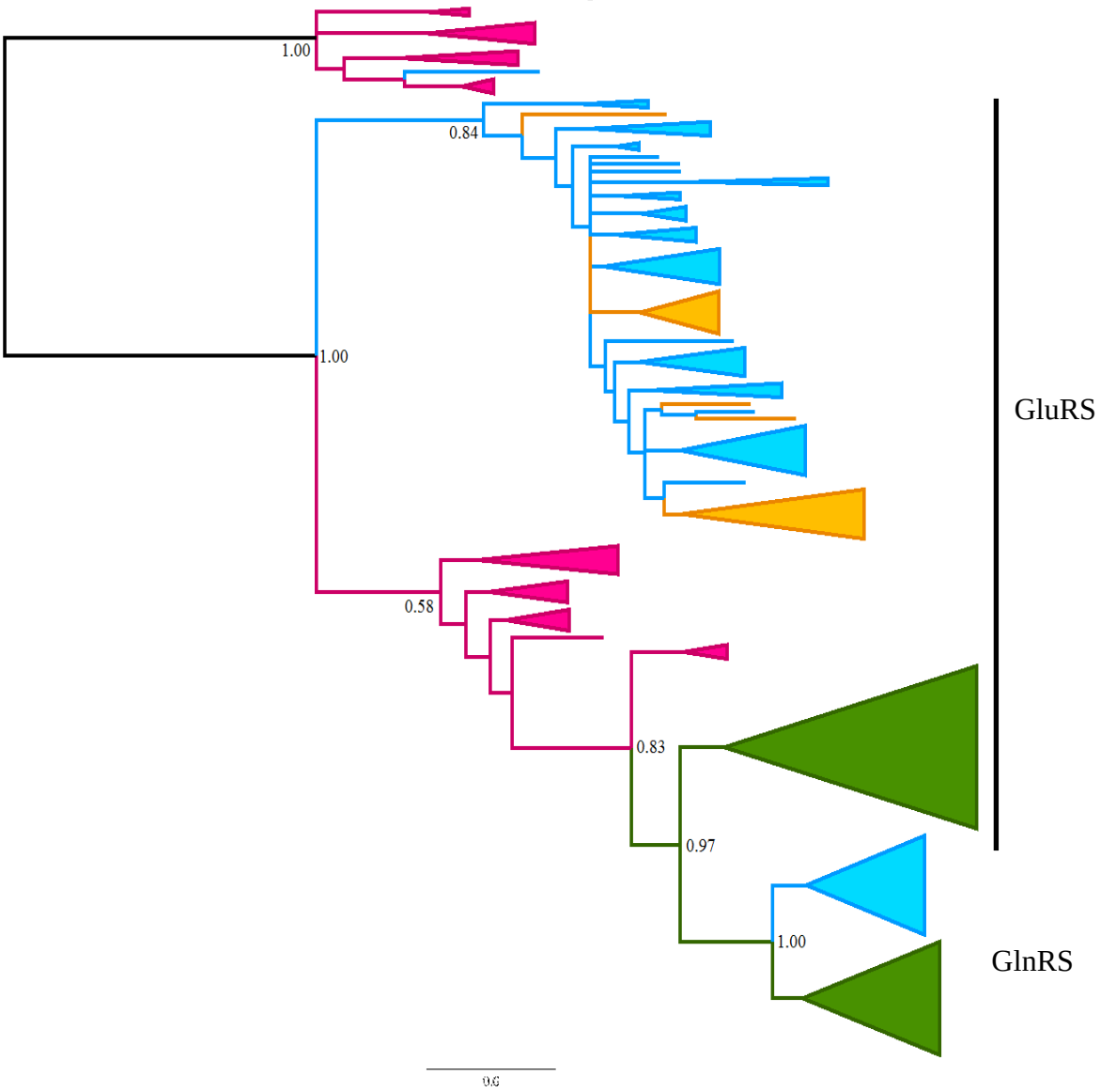


Fig. 10 Maximum likelihood composite tree of class Ic ARSs

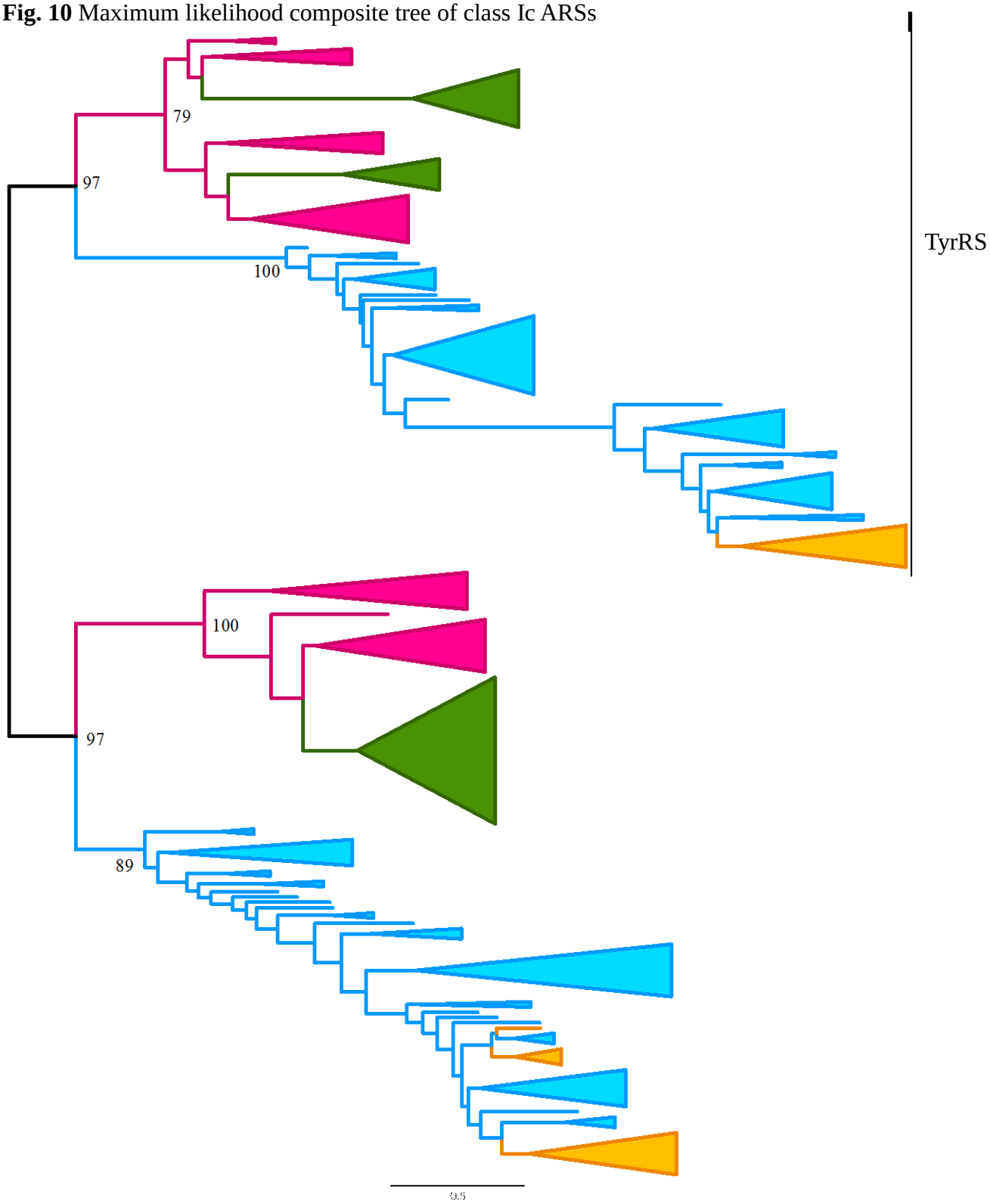


Fig. 11 Bayesian composite tree of class Ic ARSs

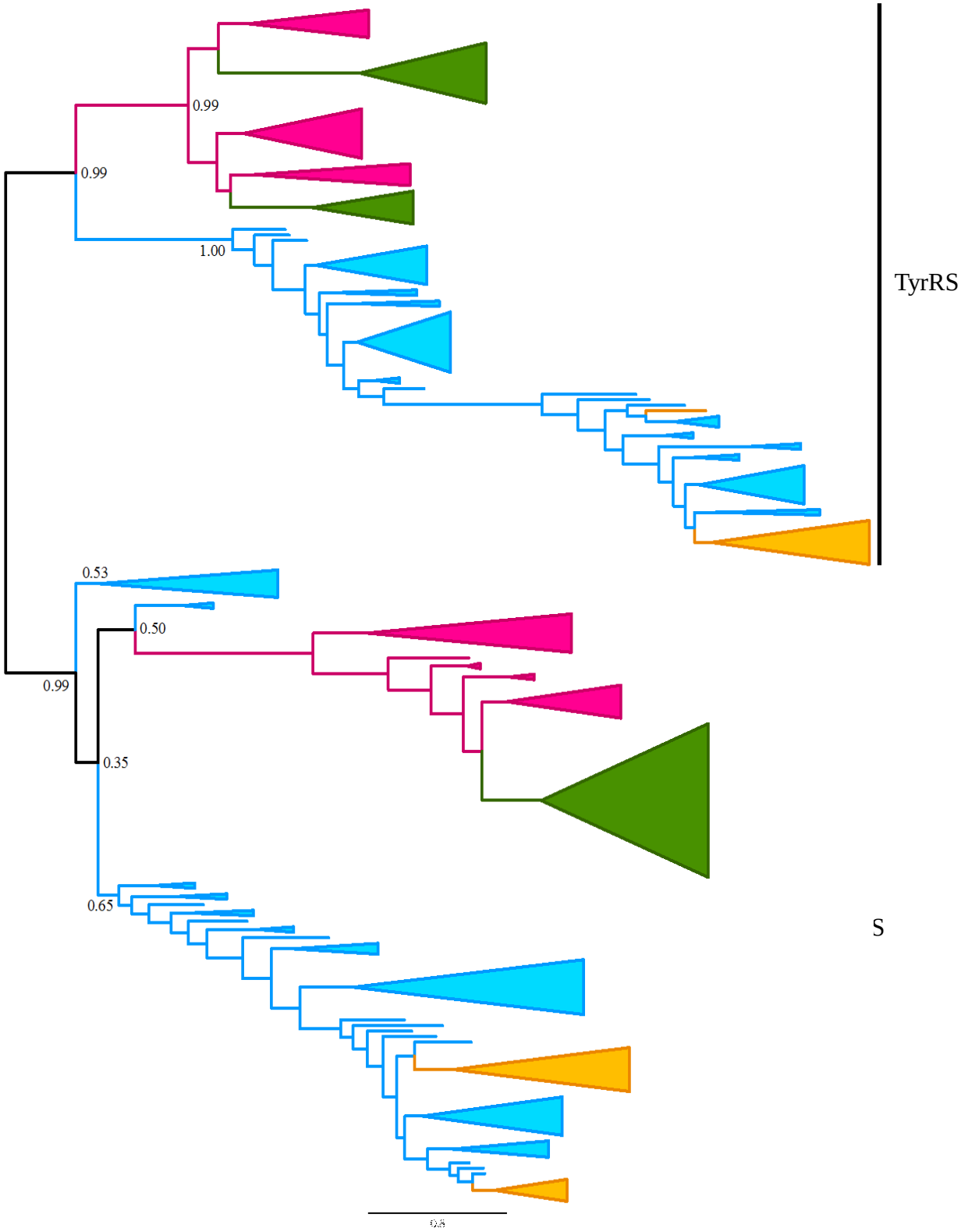


Fig. 12 Maximum likelihood composite tree of class IIa ARSs

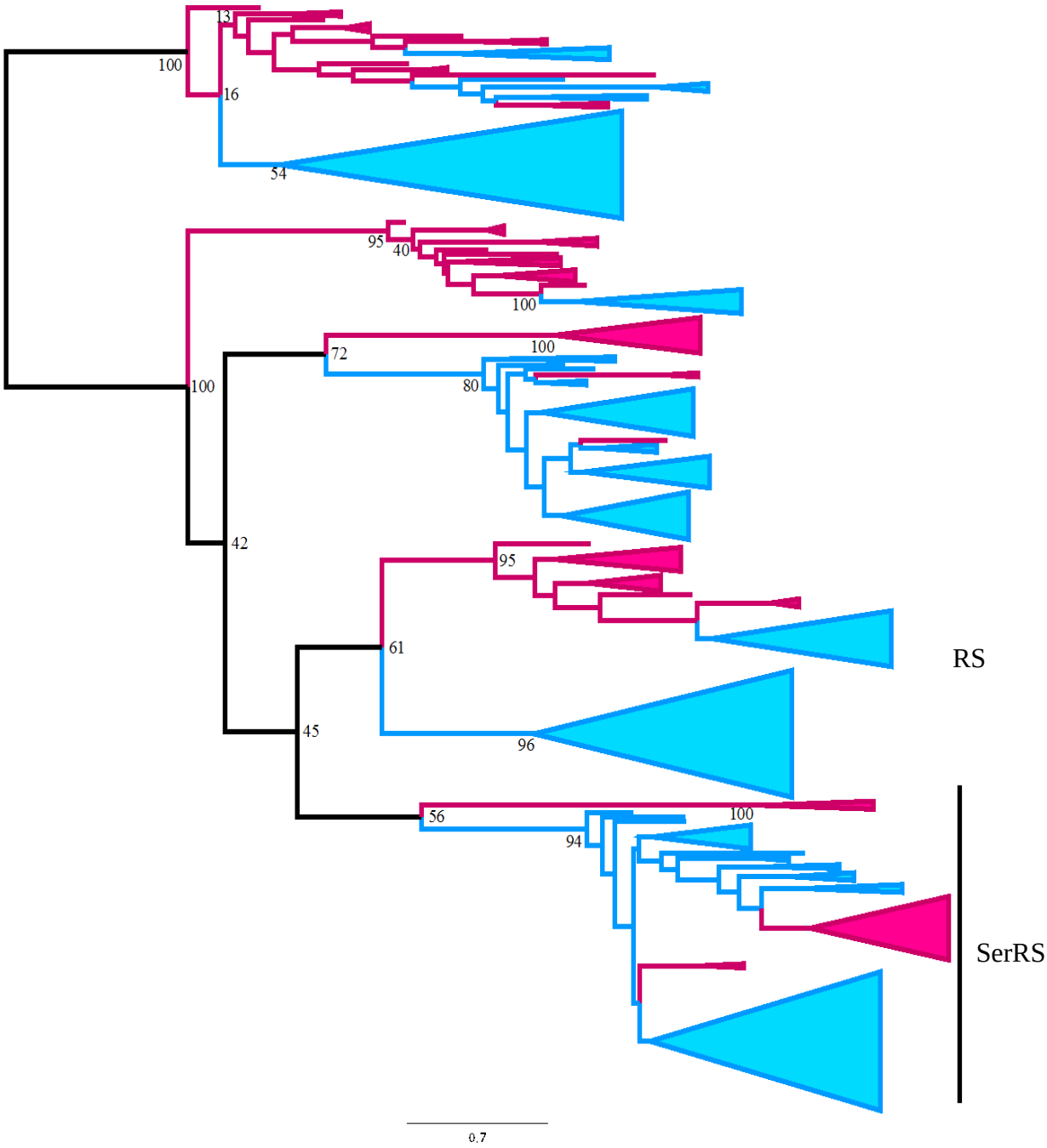


Fig. 13 Bayesian composite tree of class IIa ARSs

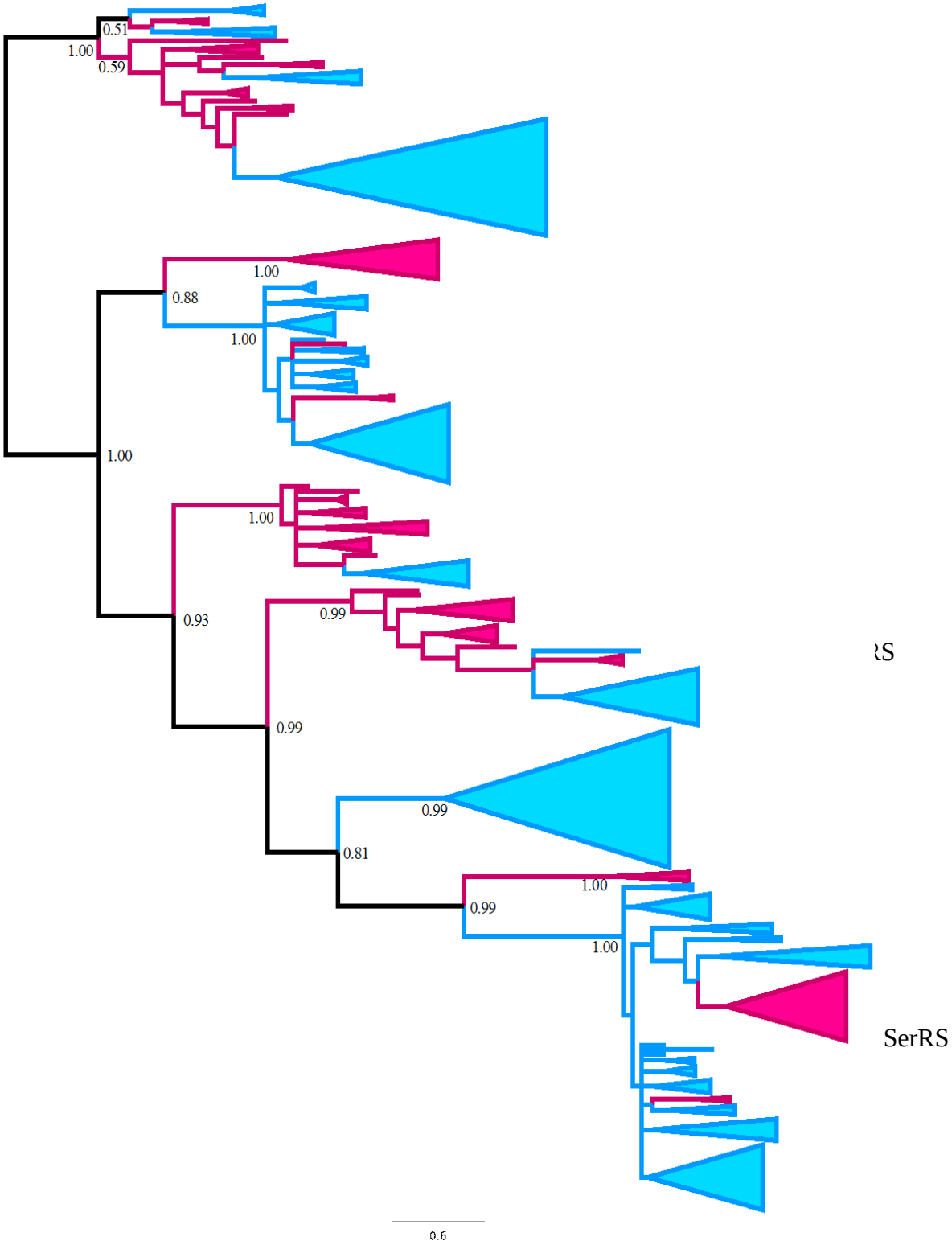


Fig. 14 Maximum likelihood composite tree of class IIb ARSs

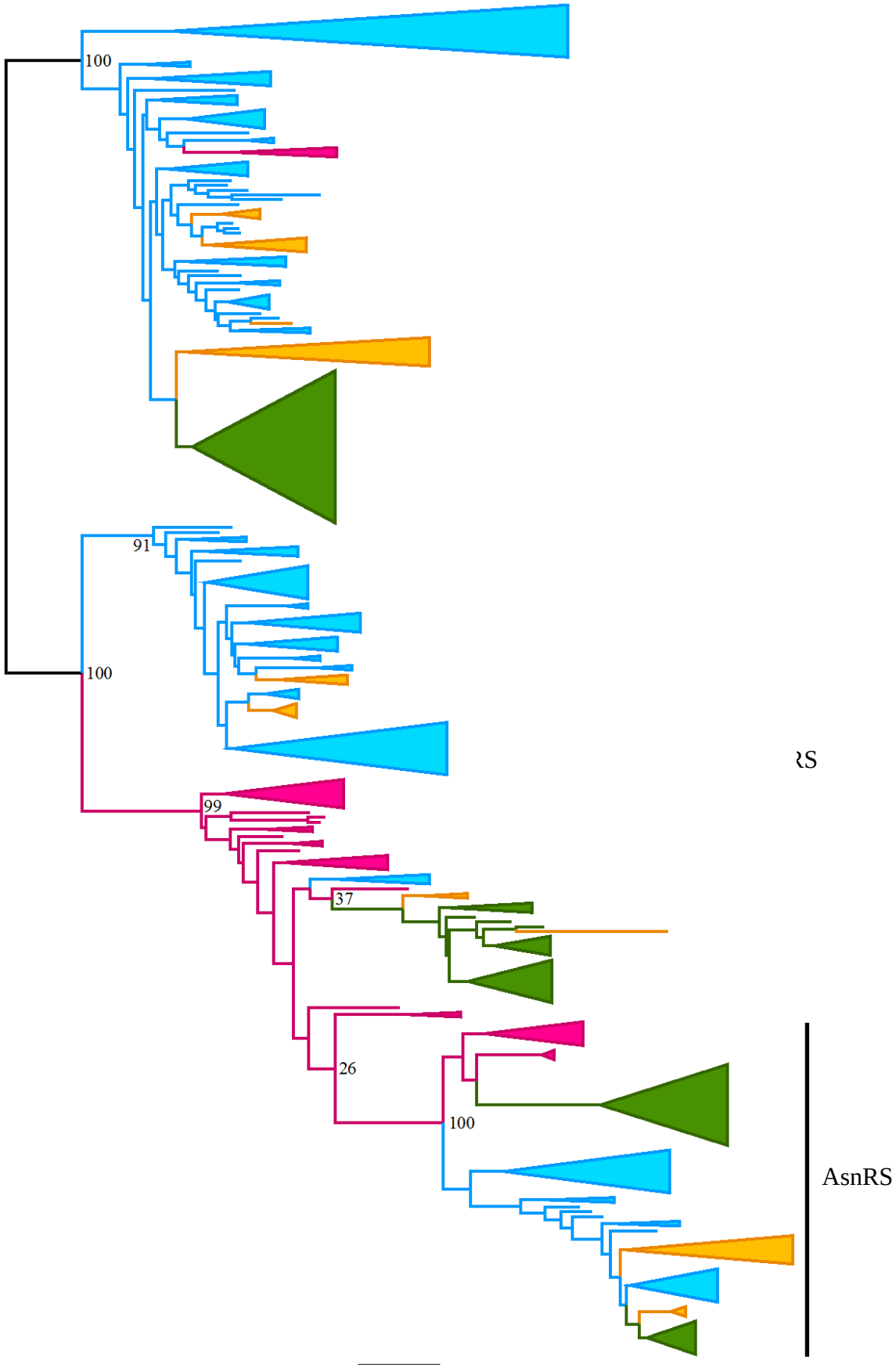


Fig. 15 Bayesian composite tree of class IIb ARSs

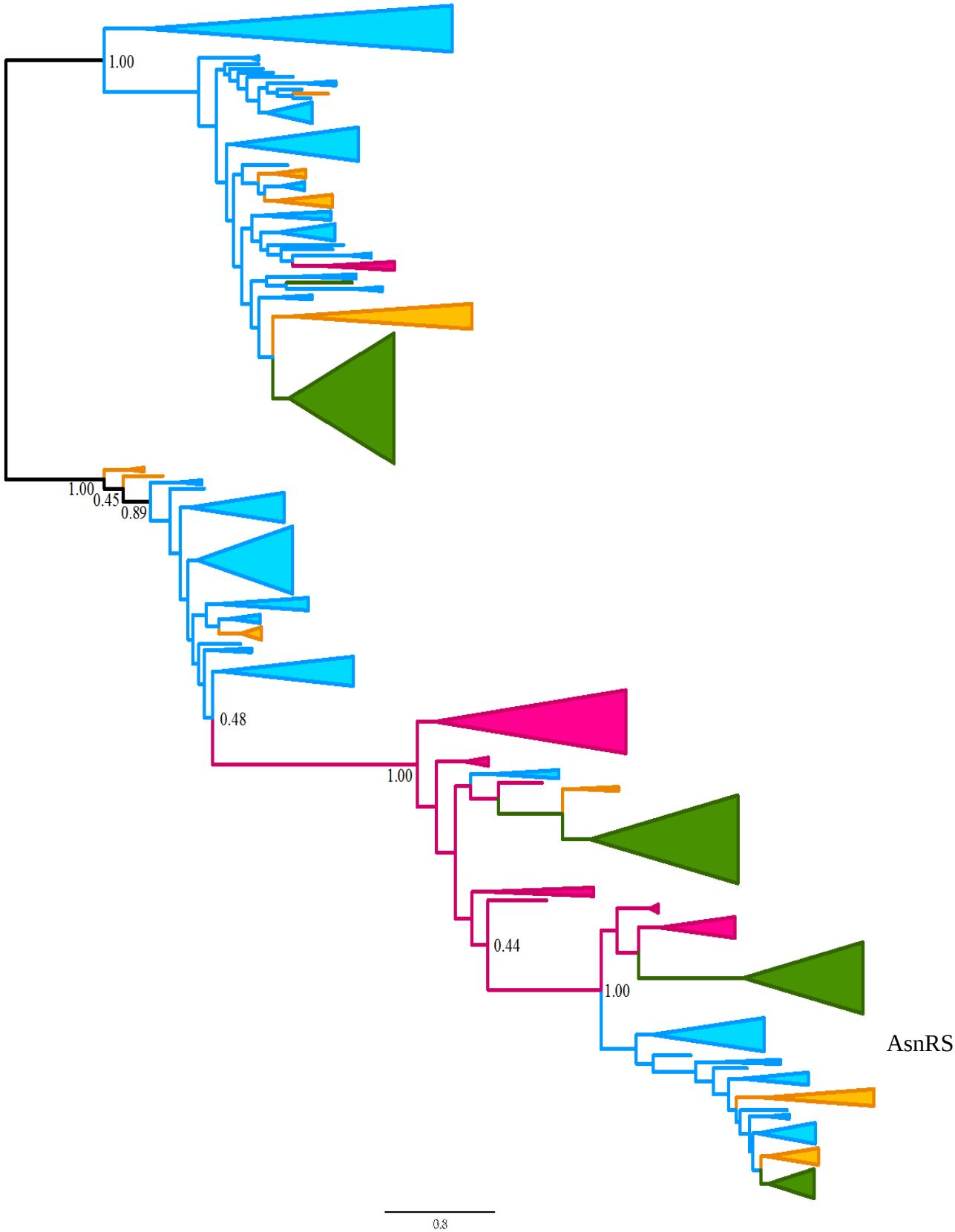


Fig. 16 Maximum likelihood composite tree of class IIc ARSs

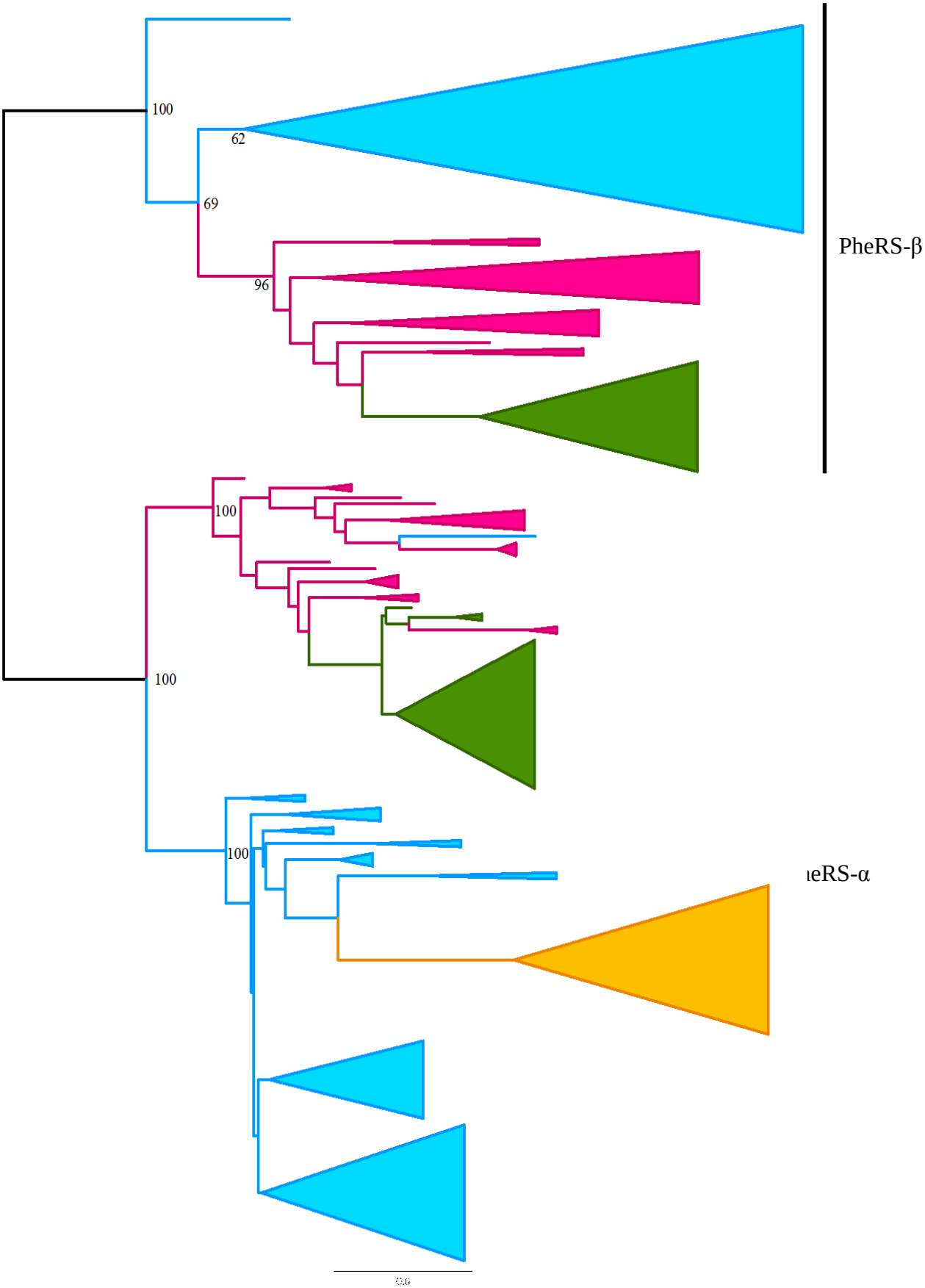


Fig. 17 Bayesian composite tree of class IIc ARSs

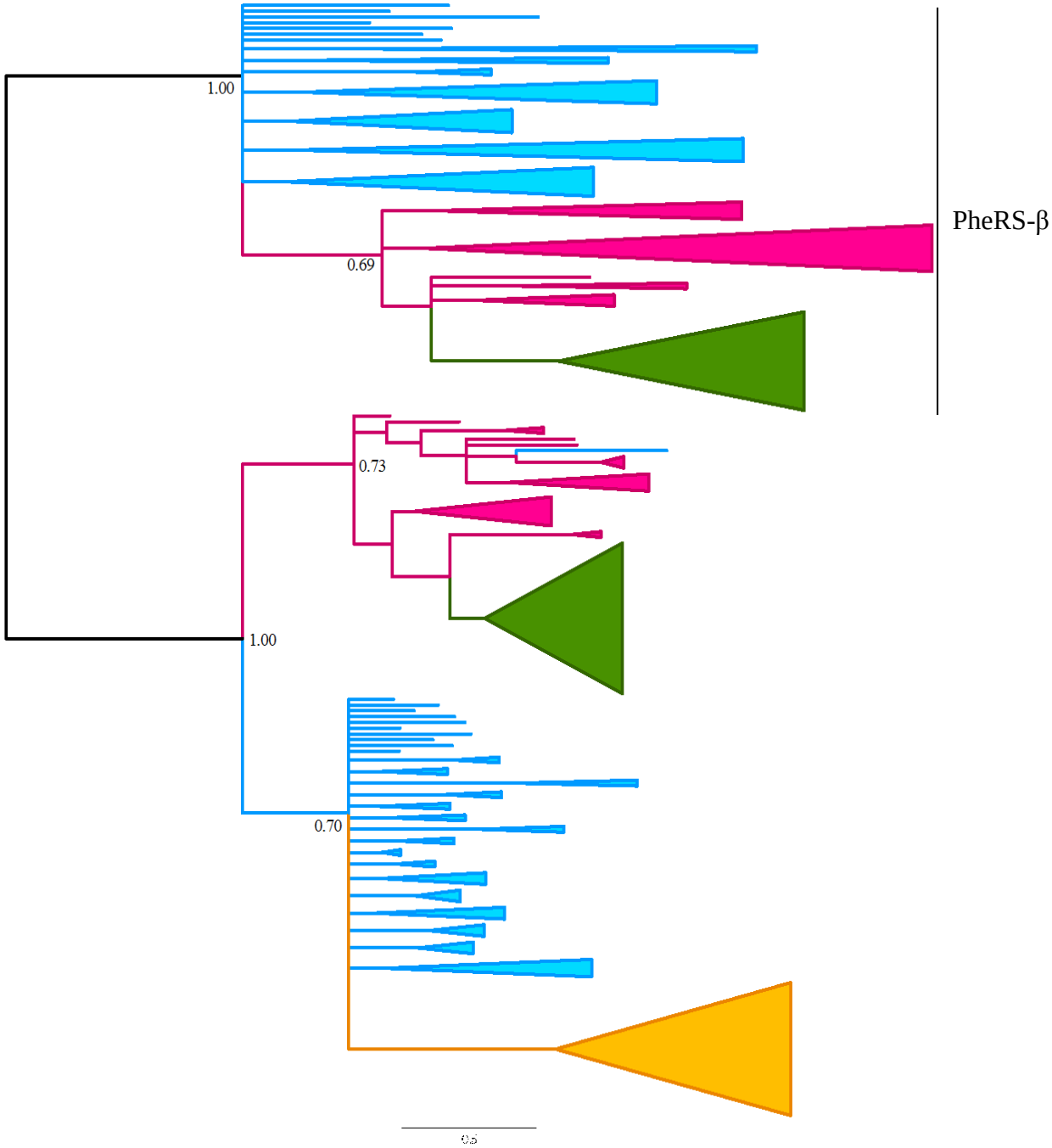


Fig. 18 Maximum likelihood composite tree of class IId ARSs

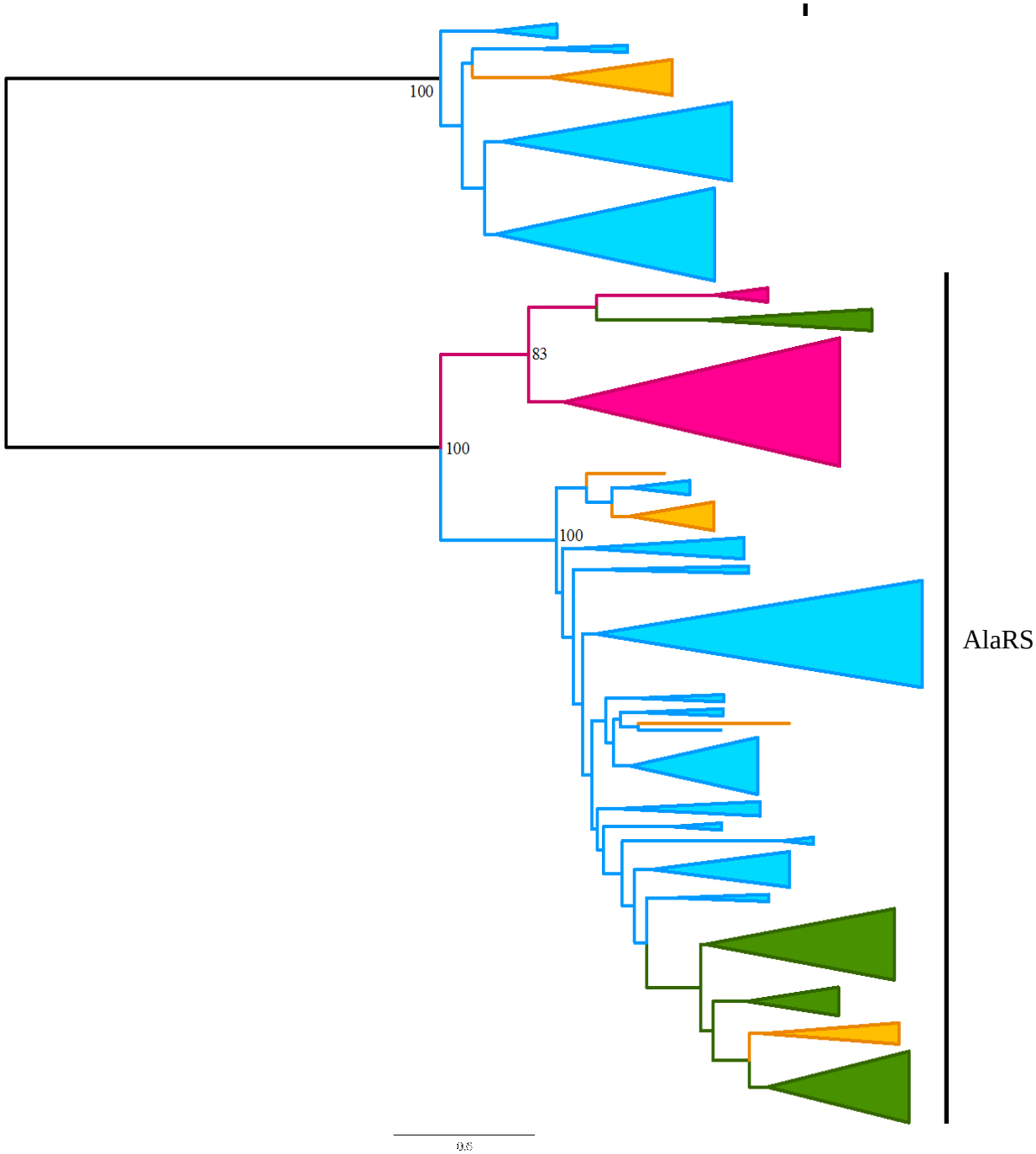
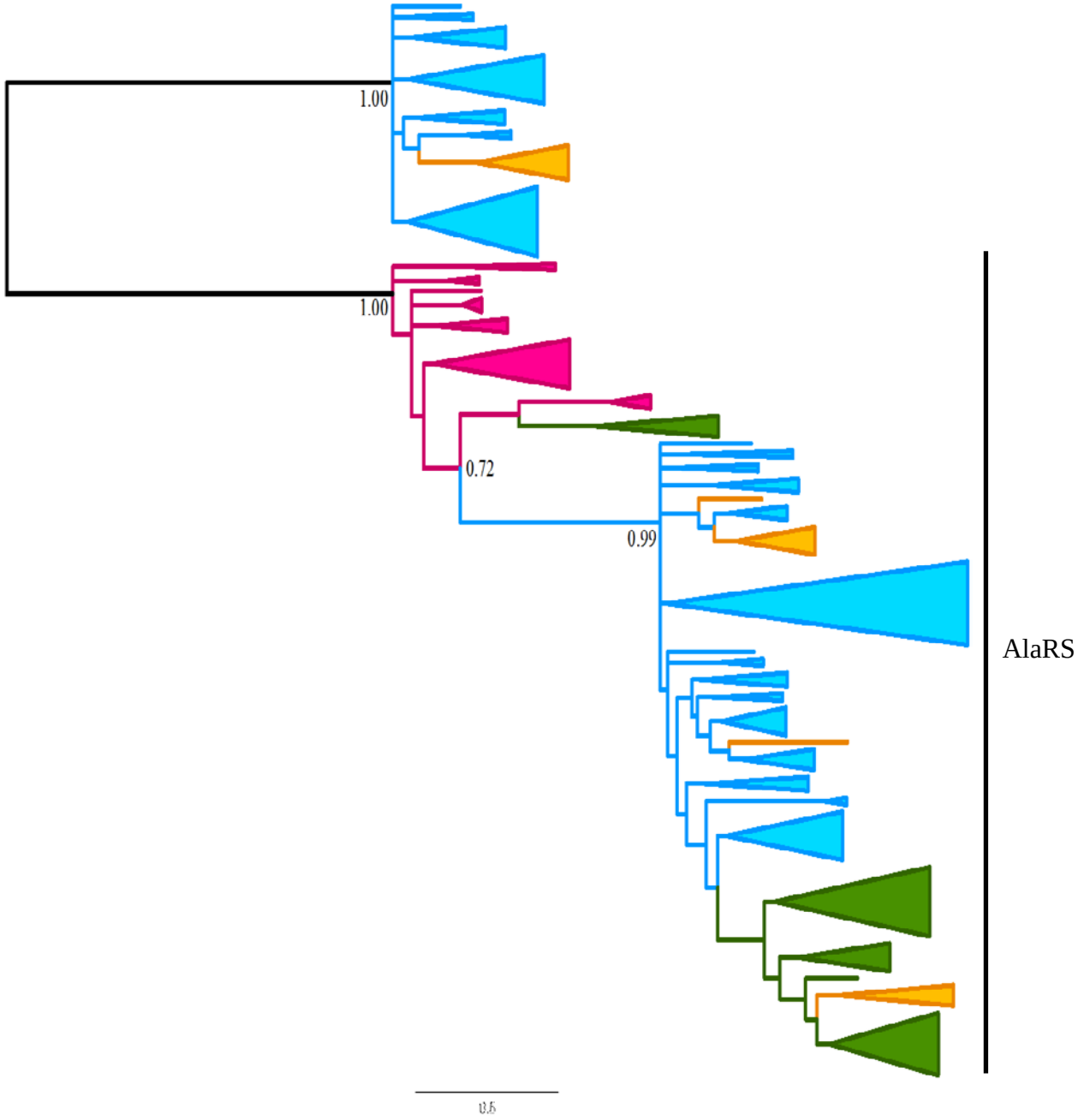


Fig. 19 Bayesian composite tree of class IId ARSs



3.7 Table

Table 5

Root placement of each ARS tree

		ML method	BI method
classIa	IleRS	(Between Bacteria and Archaea)	(Archaea)
	LeuRS	(Archaea)	(Archaea)
	ValRS	(Archaea)	(Archaea)
	MetRS	(Bacteria)	(Between Bacteria and Archaea)
	CysRS	(Archaea)	(Archaea)
	ArgRS	(Bacteria)	(Bacteria)
classIb	LysRS-class I	*Archaea*	*Archaea*
	GluRS	Between Bacteria and Archaea	Between Bacteria and Archaea
	GlnRS	Eukarya	Eukarya
classIc	TyrRS	Between Bacteria and Archaea	Between Bacteria and Archaea
	TrpRS	Between Bacteria and Archaea	(Bacteria)
classIIa	SerRS	Between Bacteria and Archaea	Between Bacteria and Archaea
	ThrRS	Between Bacteria and Archaea	Between Bacteria and Archaea
	ProRS	Between Bacteria and Archaea	(Archaea)
	GlyRS-1	*(Archaea)*	*(Archaea)*
	HisRS	(Archaea)	(Archaea)
classIIb	LysRS-class II	*Bacteria*	*Bacteria*
	AspRS	Between Bacteria and Archaea	(Bacteria)
	AsnRS	Archaea	Archaea
classIIc	PheRS-α	Between Bacteria and Archaea	Between Bacteria and Archaea
	PheRS-β	(Bacteria)	(Bacteria)
classIId	AlaRS	Between Bacteria and Archaea	(Archaea)
	GlyRS-2	*Bacteria*	*Bacteria*

Note: The root position with low resolution is in parenthesis. The root position of ARS with more than one type is indicated in italic letters. The root related to the appearance of the ARS unrelated to *C. commonote* is underlined.