Assessing the ubiquity and origins of structural maintenance of chromosomes (SMC) proteins in eukaryotes

Mari Yoshinaga[1] and Yuji Inagaki[1,2].

[1] Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan.

[2] Center for Computational Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan.

Corresponding author: Yuji Inagaki, yuji@ccs.tsukuba.ac.jp

Keywords: condensin, cohesin, chromosome assembly, chromosome segregation, DNA repair, ATPase.

ABSTRACT

Structural maintenance of chromosomes (SMC) protein complexes are common in Bacteria, Archaea, and Eukaryota. SMC proteins, together with the proteins related to SMC (SMC-related proteins), constitute a superfamily of ATPases. Bacteria/Archaea and Eukaryotes are distinctive from one another in terms of the repertory of SMC proteins. A single type of SMC protein is dimerized in the bacterial and archaeal complexes, whereas eukaryotes possess six distinct SMC subfamilies (SMC1-6), constituting three heterodimeric complexes, namely cohesin, condensin, and SMC5/6 complex. Thus, to bridge the homodimeric SMC complexes in Bacteria and Archaea to the heterodimeric SMC complexes in Eukaryota, we need to invoke multiple duplications of a SMC gene followed by functional divergence. However, to our knowledge, the evolution of the SMC proteins in Eukaryota had not been examined for more than a decade. In this study, we reexamined the ubiquity of SMC1-6 in phylogenetically diverse eukaryotes that cover the major eukaryotic taxonomic groups recognized to date (101 species in total) and provide two novel insights into the SMC evolution in eukaryotes. First, multiple secondary losses of SMC5 and SMC6 occurred in the eukaryotic evolution. Second, the SMC proteins constituting cohesin and condensin (i.e., SMC1-4), and SMC5 and SMC6 were derived from closely related but distinct ancestral proteins. Finally, we discuss how SMC1-4 were evolved from the ancestral SMC protein(s) in the very early stage of eukaryotic evolution.

INTRODUCTION

Chromosomes comprise DNA molecules, which are the body of genetic information, and a large number of proteins with diverse functions. In eukaryotes, cohesin and condensin, together with many other proteins, maintain the integrity of chromosome structure. Cohesin and condensin participate in protein complexes (Anderson et al. 2002) that bundle sister chromosomes together during mitosis (Christian H et al. 2008) and meiosis (Ishiguro 2019), and aggregate chromosomes (Sutani and Yanagida 1997), respectively. Cohesin is constituted by two Structural Maintenance of Chromosomes (SMC) proteins (SMC1 and SMC3) (Losada et al. 1998) and accessory subunits

Rad21/Scc1 and STAG1/Scc3 (Birkenbihl and Subramani 1995; Carramolino et al. 1997; Tóth et al. 1999). Condensin contains SMC2 and SMC4 (Hirano and Mitchison 1994), and a different set of accessory subunits CAP-D2, CAP-G, and CAP-H (Hirano et al. 1997). There are two additional SMC proteins, SMC5 and SMC6 (Lehmann et al. 1995; Fousteri and Lehmann 2000), which comprise the “SMC5/6” complex together with six accessory proteins (Nse1-6) (Andrews et al. 2005; Fujioka et al. 2002; Hu et al. 2005; Pebernard et al. 2004; Pebernard et al. 2006) and involve mainly in DNA repair but also replication fork stability (Aragón 2018).

SMC proteins, together with MukB, Rad50, and RecN, belong to a large ATPase superfamily with unique structural characteristics (Niki et al. 1991; Funayama et al. 1999; Löwe et al. 2001). SMC proteins comprise “head” that hydrolyzes ATP, “hinge” that facilitates the dimerization of two SMC proteins (SMC1 and SMC3 in cohesin, SMC2 and SMC4 in condensin, and SMC5 and SMC6 in the SMC5/6 complex), and antiparallel coiled coils connecting the head and hinge (Melby et al. 1998). As ATPases, SMC proteins bear seven motifs such as Walker A (P-loop), Walker B, ABC signature motif (C-loop), A-loop, D-loop, H-loop (switch motif), R-loop, and Q-loop, all of which are required ATP binding and hydrolysis. In the ATPases belonging to the SMC superfamily, the Walker A motif, A-loop, R-loop, and Q-loop are located at the N-terminus of the molecule, being remote from the rest of the motifs at the C-terminus (Palou et al. 2018). Thus, SMC proteins most likely form hairpin-like structures to make all of the sequence motifs for ATP binding in close proximity in the tertiary structures (Melby et al. 1998).

The vast majority of the members of Bacteria and Archaea possess a single SMC protein for DNA strand aggregation. In contrast to the eukaryotic SMC complexes containing heterodimeric SMC proteins, the SMC complexes in Bacteria and Archaea comprise two identical SMC proteins (i.e., homodimeric), together with accessory subunits (Britton et al. 1998; Soppa 2001). It is noteworthy that the SMC protein is not conserved strictly in Bacteria or Archaea (Soppa 2001). For instance, the absence of the conventional SMC protein in the Crenarchaeota genus *Sulfolobus* was experimentally shown to be complemented by the proteins that are distantly related to the authentic SMC, namely coalescin (Takemata et al. 2019).

To our knowledge, no study on the diversity and evolution of SMC proteins sampled from phylogenetically diverse eukaryotes has been done since Cobbe and Heck (2004). Their phylogenetic analyses recovered individual clades of SMC1-6, and further united (i) SMC1 and SMC4 clades, (ii) SMC2 and SMC3 clades, and (iii) SMC5 and SMC6 clades together. Henceforth here in this work, we designated the three unions as “SMC1+4 clan,” “SMC2+3 clan,” and “SMC5+6 clan,” respectively. Based on the phylogenies inferred from the SMC proteins in the three domains of Life, the authors proposed that SMC1-6 were yielded through gene duplication events that occurred in the early eukaryotic evolution.

The pioneering work by Cobbe and Heck (2004) was a significant first step to decipher the origin and evolution of SMC proteins in eukaryotes, albeit they provided no clear scenario explaining how a primordial SMC protein diversified into SMC1-6 prior to the divergence of the extant eukaryotes. Thus, we reassessed the ubiquity of SMC1-6 in eukaryotes and the phylogenetic relationship among the six eukaryotic SMC subfamilies in this study. Fortunately, recent advances in sequencing technology allow us to search for SMC proteins in the transcriptome and/or genome data of phylogenetically much broader eukaryotic lineages than those sampled from metazoans, fungi (including a microsporidian), land plants, and trypanosomatids analyzed in Cobbe and Heck (2004). Furthermore, computer programs for the maximum-likelihood (ML) phylogenetic methods, as well as hardwares, have been improved significantly since 2004. Thus, hundreds of SMC proteins from diverse eukaryotes can be subjected to the ML analyses now, in contrast to Cobbe and Heck (2004) wherein only a distance tree was inferred from the alignment of 148 SMC sequences.

Our survey of SMC1-6 in 101 eukaryotes confirmed the early divergence of the six SMC subfamilies in eukaryotes, albeit the secondary loss of SMC5 and SMC6 most likely has occurred in separate branches of the tree of eukaryotes. Moreover, the phylogenetic analysis of SMC1-6, bacterial and archaeal SMC, and Rad50/SbcC (304 sequences in total) disfavored the single origin of the six SMC subfamilies in eukaryotes and instead suggested that the ancestral molecule of SMC5 and SMC6 is distinct from that of SMC1-4. We finally explored multiple scenarios to explain how the repertory of the SMC subfamilies was shaped in the very early evolution of eukaryotes.

RESULTS

We first reexamined the conservation of the six SMC subfamilies in 59 phylogenetically diverse eukaryotes. Among the 59 eukaryotes, we retrieved 366 SMC proteins in total after eliminating non-SMC sequences by a preliminary phylogenetic analysis. Subsequently, we classified the 366 SMC proteins into individual subfamilies (i.e., SMC1-6) based on the phylogenetic affinity to the SMC1-6 proteins known prior to this study (Table S1). Table S1 contains truncated SMC sequences likely due to (i) the absence of the full-length transcript in the RNA-seq data and/or (ii) an intron (or introns) that hinders recovery of the entire SMC-coding region from the genome data.

Inventories of the proteins constituting cohesin and condensin

At least two of SMC1-4 were detected in 52 out of the 59 eukaryotes examined here (Fig. 1: The 7 eukaryotes, from which the set of SMC1-4 failed to be completed, are described in the latter part of this section). The ciliate Oxytricha trifallax and Paramecium tetrauelia, the parabasalid Trichomonas vaginalis appeared to have two different types of SMC4, SMC2, and SMC4, respectively (EJY74552.1 and EJY65570, for the two types of SMC4 in *Oxytricha trifallax*; XP 001443869.1 and XP 001447973 for the two types of SMC2 in *Paramecium tetrauelia*; XP_001328834.1 and XP_001328078.1 for the two types of SMC4 in *Trichomonas vaginalis*). There are also two types of SMC6 in land plants *Arabidopsis lyrata* and *Physcomitrella patens* (XP 020877425.1 and XP 020884092.1, *Arabidopsis lyrata*; XP 024388470.1 and XP_024388471.1, *Physcomitrella patens*).

From each of the telonemid *Telonema subtilis* and ancyromonad *Acryromonas sigmoides*, we detected two separate transcripts, one encoding the N-terminal portion of SMC1 and the other encoding the C-terminal portion of the same protein. Although we found no transcript encoding the entire SMC1, we regard that both *Telonema subtilis* and *Ancryromonas sigmoides* possess the standard SMC1 with the ATP binding motif split into the N and C-termini of a single molecule. Likewise, for SMC1 and SMC3 of the chrompodellid *Chromera vel*ia, SMC2 of the glaucophyte *Cyanophora paradoxa*, and SMC6 of the heterolobosean *Naegleria fowleri*, we failed to retrieve any transcript/contig encoding the entire protein in the transcriptome/genome data. Although the

uncertainties described above could not be resolved at this point, we considered that the SMC proteins mentioned above are of full-length.

We detected a subset of SMC1-4 in *Chlorokybus atmophyticus* (a green alga, Archaeplastida), *Cyanophora paradoxa*, (a glaucophyte, Archaeplastida), *Pharyngomonas kirbyi* (a heterolobosean, Discoba), *Capsaspora owczarzaki* (a filasterean, Amorphea), *Rigifila ramosa* and *Mantamonas plastica* in the CRuMs clade, and the hemimastigophoran *Hemimastix kukwesjijk* (Fig. 1). It is most likely that the 7 species described above do possess SMC1-4 but subsets of subfamilies in them were escaped from our survey due to some experimental reasons. The four SMC subfamilies were found in the phylogenetic relatives of the species mentioned above, except for *Hemimaxtix kukwesjijk* that showed no clear phylogenetic affinity to any other eukaryotes known to date (Lax et al. 2018). For instance, *Pharyngomonas kirbyi* is closely related to *Naegleria fowleri*, both of which belong to Heterolobosea, and SMC4 was not detected in the former while all of SMC1-4 were completed in the latter (Fig. 1). One of the most plausible reasons for missing one or two out of SMC1-4 in this study is the quality and quantity of sequence data. We suspect that, depending on data quality and/or quantity, all of the SMC subfamilies were difficult to detect in the species for which only transcriptome data were available. On the other hand, the SMC subfamilies detected in *Capsaspora owczarzaki* are puzzling. Both genome and transcriptome data are available for this species and we successfully identified the proteins forming cohesin, namely SMC4 and two accessory subunits investigated in this study (Rad21/Scc1 and STAG1/Scc3; see below) but SMC1. As it is difficult to imagine an “SMC1-lacking cohesin,” we most likely overlooked SMC1 for unknown reasons. In sum, despite the uncertainties above, we anticipate that all eukaryotes have the set of SMC1-4, reinforcing the significance of cohesin (including SMC1 and 3) and condensin (including SMC2 and 4) for cellular viability. The ubiquity of SMC1-4 strongly suggests that SMC1-4 were present in the last eukaryotic common ancestor (LECA).

We also investigated the conservation of accessory subunits of cohesin and condensin in the transcriptome and/or genomic data from the 59 eukaryotes (Fig. S1). In cohesin, SMC1 and SMC3 interact with Rad21/Scc1 and STAG1/Scc3. We detected Rad21/Scc1 and STAG1/Scc3 homologs in 44 and 43 out of the 59 eukaryotes

investigated here, respectively. It is worthy to note that the two accessory subunits were detected rarely in the members of Alveolata. This result implies that Rad21/Scc1 and/or STAG1/Scc3 in alveolates are highly diverged and are difficult to detect based on amino acid sequence similarity. Alternatively, some of the eukaryotes (e.g., alveolates) might have lost Rad21/Scc1 and/or STAG1/Scc3 or substitute the two proteins with as-yet-unknown proteins. Condensin comprises CAP-D2, CAP-G, and CAP-H, together with SMC2 and SMC4. CAP-D2, CAP-G, and CAP-H were found in 48, 33, and 42 out of the 59 eukaryotes, respectively. Curiously, we failed to find CAP-G in none of the members of Alveolata examined here. Likewise, neither CAP-G nor CAP-H was found in the members of Kinetoplastea examined here. Future experimental studies are necessary to examine the requirement of CAP-G in alveolates, and CAP-G and CAP-H in kinetoplastids.

Inventories of the proteins constituting the SMC5/6 complex

SMC5 and/or SMC6 were detected in 49 out of the 59 species assessed in this study (Fig. 1). A set of SMC5 and SMC6 was found in 43 out of the 49 species, while either of the two SMC proteins is missing in six species (Fig. 1). We suspected that one of the two proteins was escaped from our survey in the six species because both SMC5 and SMC6 are indispensable to constitute the SMC5/6 complex. Despite the exceptions described above, both SMC5 and SMC6 were found in phylogenetically broad eukaryotes, suggesting that the two SMC subfamilies were present in the LECA.

We could detect neither SMC5 nor SMC6 for 11 out of the 59 eukaryotes examined here (Fig. 1). Curiously, both high-quality transcriptome and genome data are available for 9 out of the 11 eukaryotes, namely three ciliates (*Paramecium tetraurelia*, *Oxytricha trifallax*, and *Tetrahymena thermophila*), and three kinetoplastids (*Trypanosoma brucei*, *Bodo saltans*, and *Perkinsela* sp.), the chrompodellid *Chromera velia*, the dinoflagellate *Symbiodinium microadriaticum*, and *Perkinsus marinus* that represents Perkinosozoa, a taxonomic group basal to dinoflagellates (Fig. 1). We here regard the aforementioned eukaryotes as the candidate lineages that discarded SMC5 and SMC6 secondarily. To pursue the possibility described above, we expanded the search on the repertoire of SMC subfamilies in (i) additional ciliates, dinoflagellates, and

chrompodellids, and (ii) additional kinetoplastids and their close relatives (i.e., diplonemids and euglenids). The results of the aforementioned surveys of the SMC subfamilies in the lineages are described in the next section. Unfortunately, we cannot be sure whether both SMC5 and SMC6 are absent truly in *Hemimaxtix kukwesjijk* and *Telonema subtilis* based on the survey solely in the transcriptome data.

The SMC5/6 complex is known to be associated with Nse1–6. In this study, we searched for the Nse1-4 sequences in the 59 eukaryotes, albeit excluded Nse5 and Nse6, both of which are little conserved among eukaryotes (Diaz and Pecinka 2018), from the survey. The survey of Nse1-4 was less successful than those of the accessory subunits in cohesin and condensin (see above). We found the set of the four Nse proteins only in 8 out of the 59 eukaryotes examined (Fig. S1). Nevertheless, at least one of Nse1-4 was found in phylogenetically diverse eukaryotes, implying that the LECA possessed the SMC5/6 complex associated with at least Nse1-4.

Secondary losses of SMC5 and SMC6

The survey of the SMC subfamilies illuminated putative secondary losses of SMC5/6 in kinetoplastids, ciliates, dinoflagellates plus *Perkinsus*, and a chrompodellid *Chromera velia* (Fig. 1). Thus, we additionally surveyed the SMC subfamilies in the transcriptome data of three kinetoplastids plus three of their relatives (two diplonemids and an euglenid), 15 ciliates, three chrompodellids, and 18 dinoflagellates. The transcriptome data of the species additionally surveyed were retrieved from public databases and assembled into the contig data. As a transcriptome analysis unlikely covers the entire repertory of the genes in the corresponding genome, all of the SMC subfamilies in the species of interest are not necessarily detectable by the survey in the contig data. Nevertheless, if SMC5/6 is truly absent in a particular group of eukaryotes, neither SMC5 nor SMC6 is unlikely detected in any members of the group of interest.

We examined whether the putative absence of SMC5/6 in kinetoplastids *Trypanosoma brucei*, *Perkinsela* sp., and *Bodo saltans* (Fig. 1) is applicable to other members of the class Kinetoplastea and members belonging to classes Diplonemea and Euglenida that comprise the phylum Euglenozoa with Kinetoplastea. As anticipated, in all of the kinetoplastids and diplonemids additionally examined, SMC5 or SMC6 was

not detected while SMC1-4 were found (Fig. 2A). In contrast, SMC1-6 were detected in the sequence data from two euglenids *Euglena gracilis* and *Eutreptiella gymnastics* (Fig. 2A). In the tree of Euglenozoa, Kinetoplastea and Diplonemea branch together by excluding Euglenida. Combining the presence/absence of SMC5/6 and the phylogenetic relationship among the three classes in Euglenozoa together, we propose that a single loss of SMC5/6 occurred in the common ancestor of kinetoplastids and diplonemids.

The results from the survey of SMC subfamilies in dinoflagellates were straightforward. Neither SMC5 nor SMC6 was found in any of the dinoflagellate species examined, whereas SMC1-4 were completed in all of them except *Togula jolla*, from which SMC1 was not found (Fig. 2B). Thus, we interpret the result as both SMC5 and SMC6 being lost in the common ancestor of dinoflagellates and perkinsids.

The presence/absence of SMC5 and SMC6 in chrompodellids is likely complicated. Importantly, both SMC5 and SMC6 were identified in *Vitrella brassicaformis* (Fig. 1). Among the species additionally examined, SMC6 was detected in *Alphamonas edax* and *Colpodella angusta* (Fig. 2B). Thus, the ancestral chrompodellid species likely possessed SMC5 and SMC6 and the loss of the two SMC subfamilies occurred at least on the branch leading to *Chromera velia*. We are currently unsure whether an undescribed chrompodellid “symbiont X” truly lacks SMC5/6, as the transcriptome was prepared from a limited number of the parasite/symbiont cells manually isolated from the host animal (Janouškovec et al. 2019). Thus, the library for sequencing may have missed the transcripts encoding SMC5 and SMC6.

Overall, the qualities/quantities of the ciliate transcriptome data may not be high enough to identify all of the six SMC subfamilies. We succeeded identifying a set of SMC1-4 was identified in only four out of the 15 species examined additionally (Fig. 2B). Nevertheless, we could find SMC5 or SMC6 in none of the 15 species examined additionally. Based on the surveys of SMC5/6 in the transcriptome data together with those in the three high-quality genome data, we propose that the common ancestor of ciliates was lost both SMC5 and SMC6 secondarily.

Phylogenetic relationship among the SMC subfamilies in eukaryotes

We subjected only SMC sequences that cover both highly conserved “block” at the N-terminus (i.e., Walker A and Q-loop motifs) to those at the C-terminus (i.e., Signature and Walker B motifs) to the ML phylogenetic analyses, along with the bacterial and archaeal SMC sequences, and Rad50/SbcC sequences (Rad50 in Bacteria is termed as SbcC). In the ML analysis, the six SMC subfamilies in eukaryotes formed individual clades with MLBPs ranging from 83% to 100% (Fig. 3A). The clade of the bacterial SMC and that of Rad50/SbcC were supported by MLBPs of 97% and 95%, respectively, albeit the monophyly of the archaeal SMC received only an MLBP of 41%.

In the SMC phylogeny presented in Cobbe and Heck (2004), SMC1 and SMC4, SMC2 and SMC3, and SMC5 and SMC6 formed the SMC1+4 clan, the SMC2+3 clan, and the SMC5+6 clan, respectively, although no bootstrap analysis was carried out. In contrast, we subjected the alignment containing SMC and Rad50/SbcC in Bacteria, Archaea, and Eukaryota to both ML and ML bootstrap analyses, and the SMC1+4 clan, SMC2+3 clan, and SMC5+6 clan recovered with MLBPs of 94%, 53%, and 86%, respectively (Fig. 3A). The bacterial and archaeal SMC formed the “prokaryotic SMC clan” with an MLBP of 51% (Fig. 3A). Significantly, the Rad50/SbcC clade (outgroup) attached to the branch leading to the SMC5+6 clan. According to the rooted ML tree topology, the SMC5+6 clan and then the SMC2+3 clan branched in this order prior to the separation of the SMC1+4 clan and prokaryotic SMC clan. Thus, the ML tree topology disfavored the monophyly of the six SMC subfamilies in eukaryotes (Figs. 3A & 3B). The MLBP for the earliest branch of the SMC5+6 clan was 77%, and the SMC1+4 clan and prokaryotic SMC clan grouped together with an MLBP of 58%. As the nodes connecting the four SMC clans were supported only by moderate MLBPs, we should be cautious not to accept the tree topology inferred by the ML method as it is and need to explore alternative scenarios for the evolution of the SMC subfamilies in eukaryotes. Indeed, the alternative tree, which was generated by modifying the ML tree topology to connect SMC1+4 clan directly to SMC2+3 clan, failed to be rejected in an AU test (P = 0.0682; Tree 2 in Fig. 3B). On the other hand, another alternative tree, in which the monophyly of SMC1-6 was enforced, was rejected at a 1% level (P = 0.00726; Tree 3 in Fig. 3B). The results described above consistently suggested that the SMC5+6 clan is

evolutionarily distantly related from the rest of the SMC subfamilies in eukaryotes (i.e., SMC1-4).

DISCUSSION

LECA possessed a set of cohesin, condensin, and the SMC5/6 complex

The pioneering work by Cobbe and Heck (2004) proposed the ubiquity of SMC1-6 among eukaryotes, albeit their conclusion was drawn from the survey against a phylogenetically restricted set of eukaryotes (i.e., 16 opisthokonts, two land plants, and two trypanosomatids). Transcriptome and/or genome data are now available for phylogenetically diverse eukaryotes so that the current study could search SMC proteins in a much broader diversity of eukaryotes than Cobbe and Heck (2004). We subjected 59 species, which represent all of the currently recognized major branches in the tree of eukaryotes plus “orphan” species. Fortunately, the search conducted in this study confirmed the ubiquity of SMC1-6 among eukaryotes (Fig. 1). We firmly conclude that the six SMC subfamilies had already existed in the genome of the LECA and the LECA most likely used cohesin, condensin, and the SMC5//6 complex for the maintenance of chromosomes.

Besides SMC1-6, non-SMC proteins (accessory subunits) comprise cohesin, condensin, and the SMC5//6 complex but their phylogenetic distributions have not been examined prior to this study. We searched for Rad21/Scc1 and STAG1/Scc3 in cohesin, CP-D2, CAP-G, and CAP-H in condensin, and Nse-1-4 in the SMC5/6 complex, in the 59 species in this study (Fig. S1). Due to the relatively low conservation at the amino acid sequence level, the search of the accessory subunits was less successful than that of the SMC proteins with highly conserved sequence motifs for ATP-binding. Nevertheless, we found the accessory subunits across the tree of eukaryotes, suggesting that the subunit compositions of cohesin, condensin, and the SMC5/6 complex of the LECA were similar to those of modern eukaryotes.

Although the early emergence of cohesin, condensin, and the SMC5/6 complex in the eukaryotic evolution (see above), we unveiled the secondary losses of SMC5 and SMC6 on separate branches in the tree of eukaryotes. By combining the survey of SMC1-6 in 59 phylogenetically diverse eukaryotes plus additional 42 species, we

propose that secondary loss of SMC5 and SMC6 occurred in (i) the common ancestor of kinetoplastids and diplonemids, (ii) that of dinoflagellates and perkinsids, (iii) that of ciliates, and (iv) on the branch leading to the chrompodellid *Chromera velia* (Figs. 2A &2B). Importantly, our results imply that the SMC5/6 complex is not absolutely indispensable for cell viability. If so, extra eukaryotic groups/species lacking SMC5 and SMC6 likely have been overlooked among those of which large-scale sequence data are currently not available, and novel ones that are currently unknown to science.

SMC5 and SMC6 are not the direct descendants of the SMC proteins in Bacteria and Archaea

The majority of bacteria and archaea possess a single type of SMC, suggesting that the two identical SMC proteins form a homodimeric complex (Soppa 2001). In contrast, eukaryotes possess the six SMC subfamilies to constitute the three distinct heterodimeric complexes, with exception of some groups lacking SMC5 and SMC6 (see above). It has been assumed that the SMC subfamilies in eukaryotes emerged from a single ancestral SMC through multiple rounds of gene duplication followed by functional divergence (Cobbe and Heck 2000; Cobbe and Heck 2004). To retrace how the repertory of the SMC subfamilies was shaped in the LECA, the phylogenetic relationship of the SMC subfamilies in Bacteria, Archaea, and Eukaryota is critical.

To phylogenetically examine the evolution of SMC in Bacteria, Archaea, and Eukaryota, SMC-related proteins (e.g., Rad50/SbcC), which had been present prior to the three domains, are indispensable as the outgroup. If all of the six SMC subfamilies in eukaryotes emerged from a single ancestral SMC protein, the monophyly of SMC1-6 should be recovered by excluding the bacterial and archaeal homologs in the rooted SMC phylogeny. Curiously, as shown in Fig. 3A, the ML analyses placed the prokaryotic SMC clan within the eukaryotic SMC subfamilies and the root of the SMC phylogeny located on the branch leading to the SMC5+6 clan. It is worthy to note that Cobbe and Heck (2004) presented the essentially same tree topology inferred from the alignment of SMC plus SMC-related proteins [Fig. 5 in Cobbe and Heck (2004)]. The MLBP value for the grouping SMC1-4, the bacterial, and archaeal SMC proteins appeared to be not conclusive (MLBP of 77%; Fig. 3A), albeit an AU test rejected the alternative tree with

the monophyly of SMC1-6 (Fig. 3B). Thus, we conclude that SMC5 and SMC6 are distantly related to SMC1-4, suggesting that the ancestral molecule of SMC5 and SMC6 was distinct from that of SMC1-4. To our knowledge, there is no bacterial or archaeal homolog of SMC5 or SMC6, suggesting that the two particular SMC proteins are of eukaryote-specific.

SMC evolution before and after the split between Archaea and Eukaryotes

In this section, we discuss the scenarios explaining how SMC1-4 had emerged in the early stage of the eukaryotic evolution based on two possible phylogenetic relationships between SMC1+4 and SMC2+3 clans relative to the prokaryotic SMC clan. First, we examine the evolution of SMC1-4 assuming the tree topology in which the SMC1+4 and prokaryotic SMC clans were tied together excluding the SMC2+3 clan. This tree topology corresponds to the rooted SMC phylogeny (Fig. 3A).

Combining the rooted SMC phylogeny (Fig. 3A) and the widely accepted phylogenetic affinity between Archaea (Asgard archaea in particular) and Eukaryota together, it is straightforward to assume that SMC1 and SMC4 are the direct descendants of the authentic SMC in Archaea. In addition, we need to hypothesize an as-yet-undescribed protein, “SMCX,” prior to the split between Archaea and Eukaryota (Fig. 4A). SMCX needs to fulfill the two conditions described below. First, SMCX is closely related to but distinct from the authentic SMC in Bacteria and Archaea. Second, SMCX is ancestral to SMC2 and SMC3 in the LECA and their descendants (i.e., modern eukaryotes). Importantly, there is no sign of SMCX in any currently known archaea including Asgard archaea, which are currently the most plausible candidate for the closest relative of eukaryotes (Zaremba-Niedzwiedzka et al. 2017; Williams et al. 2020). Thus, we further need to invoke the secondary loss of SMCX in the archaeon from which the first eukaryote emerged (designated as “archaeon*”; Fig. 4A). We cannot exclude the possibility of the archaeon*, which is closer to eukaryotes than Asgard archaea, being overlooked in nature. If so, this hypothetical cell might have retained both authentic SMC and SMCX. Finally, to bridge archaeon* with two SMC proteins and the LECA with SMC1-4, we propose that the “primordial eukaryote,” which existed prior to the LECA, used two distinct SMC proteins, one derived from the

authentic SMC and is ancestral to SMC1 and SMC4 (designated here as “SMC¼”), and the other derived from SMCX and is ancestral to SMC2 and SMC3 (designated here as “SMC⅔”) (Fig. 4A). SMC¼ and SMC⅔ likely constituted a heterodimeric SMC complex in the premordial eukaryote.

In the scenario discussed above, we incorporated no lateral gene transfer (LGT) to the evolution from SMCX to SMC2 and SMC3. Nevertheless, if we allow LGT in the SMC evolution, the origin of SMCX is not necessary to be archaeon* (Fig. 4B). In the scenario invoking LGT, SMC¼ in the primordial eukaryote was derived from the authentic SMC inherited vertically from archaeon*, whereas SMC⅔ could have been derived from SMCX in a bacterium or archaeon that was distantly related to archaeon* (Fig. 4B). MukB and coalescin, which have been found in γ-proteobacteria (and several other bacteria) and crenarchaeotes *Sulfolobus* spp., respectively, have been known as the functionally homologous to the authentic SMC. Nevertheless, the amino acid sequences of MukB and coalescin are extremely divergent from that of the authentic SMC, suggesting that neither MukB nor coalescin can be the candidate for SMCX that gave rise to SMC2 and SMC3 in eukaryotes. Once SMC¼ and SMC⅔ were established in the primordial eukaryote, the later SMC evolution in eukaryotes was most likely driven by two independent but concomitant gene duplications followed by functional divergence (highlighted by diamonds in Figs. 4A & B).

The rooted SMC phylogeny tied the SMC1+4 clan with the prokaryotic SMC clan instead of the SMC2+3 clan (Fig. 3A). Nevertheless, the possibility of the direct union of SMC1+4 and SMC2+3 clans failed to be excluded (Fig. 3B), leaving room for the single origin of SMC1-4. In the scenario assuming SMC1-4 being derived from a single SMC protein (Fig. 4C), archaeon* possessed the authentic SMC alone as do the currently known members in Archaea. In the primordial eukaryote, a single gene duplication followed by functional divergence yielded SMC¼ and SMC⅔ from the authentic SMC (Fig. 4C). In the very early eukaryotic evolution from the primordial eukaryote to LECA, the four SMC subfamilies diverged from SMC¼ and SMC⅔ through separate gene duplications followed by functional divergence.

CONCLUSIONS

In this study, we reexamined the ubiquity and evolution of the six eukaryotic SMC subfamilies based on the transcriptome and genome data accumulated from phylogenetically diverse eukaryotes. The survey of SMC sequences conducted in this study confirmed the ubiquity of SMC1-6 in eukaryotes (Fig. 1). Significantly, we here revealed two novel aspects of the SMC evolution in eukaryotes. First, we identified multiple secondary losses of SMC5 and SMC6 in the tree of eukaryotes (Figs. 2A & 2B), questioning their absolute necessity for cell viability. Second, we noticed that SMC5 and SMC6, and SMC1-4 were most likely derived from distinct ancestral molecules. The rooted SMC phylogeny suggests that SMC1-4 share the ancestry with the bacterial and archaeal SMC, albeit no bacterial or archaeal protein bearing a special phylogenetic affinity to SMC5 and SMC6 has been known to date (Figs. 3A & 3B). Finally, we discussed the scenarios for the evolution of SMC1-4 based on two tree topologies, one in which SMC1-4 group together and exclude the bacterial and archaeal SMC and the other in which the bacterial and archaeal SMC were nested within SMC1-4 (Fig. 3B). The scenario assuming the monophyly of SMC1-4 (Fig. 4C) is much simpler than the other in which a hypothetical molecule (SMCX) needs to be invoked (Figs. 4A & 4B). We favor the simpler scenario, albeit the monophyly of SMC1-4 failed to be recovered in the rooted SMC phylogeny (Fig. 2A). Thus, it is currently difficult to draw the definite conclusion on the evolution of SMC1-4.

MATERIALS AND METHODS

Surveys of SMC proteins

To examine the ubiquity of SMC1-6 in eukaryotes, we searched for SMC amino acid sequences in 59 phylogenetically diverse species. For 36 out of the 59 eukaryotes, the SMC sequences were searched for in NCBI GenBank non-redundant (nr) protein sequences database by BLASTP (Camacho et al. 2009). For the rest of the 23 eukaryotes, the SMC sequences were retrieved from the corresponding transcriptome data. The searches against the GenBank nr database and those against the contig data assembled from the transcriptome data are separately explained below.

For each of the 36 species, a BLASTP search was conducted by using the amino acid sequences of human SMC homologs as queries (GenBank accession numbers are Q14683 for SMC1, O95347 for SMC2, Q9UQE7 for SMC3, Q9NTJ3 for SMC4, Q8IY18 for SMC5, and Q96SB8 for SMC6). The amino acid sequences, which showed high similarity to the human SMC1-6 in the initial BLAST analyses, were subjected to reciprocal BLASTP against the GenBank nr protein database. We retained the sequences, which showed the similarities to the SMC sequences deposited in the database in the second BLAST analyses, as the candidates for the SMC sequences.

For the 23 eukaryotes, we downloaded their transcriptome data from NCBI Sequence Read Archive (Sayers et al. 2020) to identify the transcripts encoding SMC proteins. The raw sequence reads were trimmed by fastp v0.20.0 (Chen et al. 2018) with the -q 20 -u 80 option and then assembled by Trinity v2.8.5 or v2.10.0 (Grabherr et al. 2011). The candidate sequences of SMC proteins were retrieved from the assembled data by TBLASTN with the human SMC1-6 as queries followed by reciprocal BLASTP against the GenBank nr protein database.

The search of SMC proteins in the 59 eukaryotes revealed the secondary loss of SMC5 and SMC6 in phylogenetically distinct taxonomic groups, namely kinetoplastids, dinoflagellates plus *Perkinsus marina*, ciliates, and *Chromera velia* (see the Results). Thus, we additionally assess the repertories of SMC1-6 in the phylogenetic relatives of the above-mentioned species (42 species in total). For most of the 42 species, only have transcriptomic data available and thus the survey of SMC sequences against the transcriptome data was repeated as described above.

For 17 bacteria and 14 archaea, the SMC sequences were separately searched for in the GenBank nr protein database with the *Bacillus subtilis* SMC (GenBank accession number NC_000964.3) and *Archaeoglobus fulgidus* SMC (NC_000917.1) as queries, respectively. The detailed procedure of the search in the GenBank database was the same as described above.

Surveys of Rad50/SbcC proteins

Several proteins related to SMC (SMC-related proteins) have been known in Bacteria, Archaea, and/or Eukaryota. For instance, Rad50 is one of the SMC-related proteins

conserved among the three domains of life. We retrieved Rad50 amino acid sequences of 15 eukaryotes and 20 archaea from the NCBI GenBank nr database by the BLAST search as described above. The BLAST searches for eukaryotic and archaeal Rad50 were separately conducted with the human (GenBank accession number Q92878) and *A. fulgidus* Rad50 (NC_000917.1) as queries, respectively. The amino acid sequences of SbcC of 16 bacteria were retrieved from the GenBank database through keyword searches.

Phylogenetic analyses

We newly identified the candidate sequences for SMC proteins sampled from phylogenetically broad eukaryotes (see above). The SMC candidate sequences identified in this study and SMC sequences, which have been known for extensively studied model eukaryotes (e.g., animals, fungi, and land plants), were subjected to the maximum-likelihood (ML) phylogenetic analysis with the LG + Γ + I model. This preliminary analysis aimed to filter extremely divergent sequences and the sequences that could not be classified into any of the eukaryotic SMC sub-families with confidence. We examined the ubiquity of each of the six SMC subfamilies in eukaryotes based on the filtered SMC sequences (298 sequences in total).

To retrace the evolution of the six SMC subfamilies in eukaryotes, we further selected 240 SMC sequences containing both of the two ATP-binding motifs in the N- and C-termini from the 298 eukaryotic SMC sequences. After the exclusion of redundant sequences by CD-HIT (Fu et al. 2012), we retained 222 eukaryotic SMC sequences for the phylogenetic analyses described below. Individual sub-families of eukaryotic SMC were separately aligned by MAFFT v7.453 (Katoh and Standley 2013) with L-INS-I option. Likewise, 17 bacterial SMC, 14 archaeal SMC, and 51 Rad50/SbcC amino acid sequences were separately aligned as described above. The alignments of SMC1-6, bacterial SMC, archaeal SMC, and Rad50/SbcC were combined into the “SMC/SMC-related” alignment by MAFFT with --merge option. The resultant alignment contained 37 of SMC1, 41 of SMC2, 37 of SMC3, 37 of SMC4, 34 of SMC5, and 36 of SMC6 sequences. Manual exclusion of ambiguously aligned positions, followed by trimming of gap-containing positions by trimAI v1.2 (Capella-Gutiérrez et al. 2009) with -

gt 0.8 option, left 290 unambiguously aligned amino acid positions in 253 SMC and 51 Rad50/SbcC for the phylogenetic analyses described below.

We subjected the SMC/SMC-related alignment to the ML analysis with the LG + Γ + F + C60 + PMSF model. The guide tree was reconstructed with the ML method with the LG + Γ + I model. The reliability of each bipartition recovered in the ML tree was examined by a 100-replicate non-parametric ML bootstrap analysis. IQ-TREE v1.6.12 (Nguyen et al. 2015) was used for all of the ML phylogenetic analyses conducted in this study. The SMC/SMC-related alignment was also analyzed with Bayesian method by PhyloBayes v4.1 (Lartillot et al. 2009) using the CAT+GTR model. We ran four Markov chain Monte Carlo (MCMC) chains more than 1,000,000 cycles but the indicator of the convergence among the MCMC chains (maxdiff) remained large. Thus, we regard that the tree inferred from the PhyloBayes analysis is not suitable to discuss the phylogenetic relationship among the eukaryotic SMC subfamilies.

SMC/SMC-related alignments and the amino acid sequences included in the alignment are available at https://drive.google.com/drive/folders/1w9UA7dZJnYdZS566aFYODOmAf5q5OPT9?usp=sharing.

Surveys of accessory subunits of the SMC complexes in eukaryotes

Cohesin contains two accessory subunits Rad21/SCC1 and STAG1/SCC3, together with SMC1 and 3, while SMC2 and 4 interact with CAP-H, CP-D2, and CAP-G to constitute condensin. Nse1, Nse2, Nse3, Nse4, Nse5, and Nse6 are known to participate in the SMC5/6 complex. We searched for the accessory subunits in cohesin, condensin, and the SMC5/6 complex in the 59 eukaryotes. For the subunits of the SMC5/6 complex, only Nse1-4 were searched for in this study, because Nse5 and Nse6 were little conserved sequence amongst eukaryotes (Diaz & Pecinka, 2018). The BLAST searches of the aforementioned proteins were conducted as described above. The human homologs are used as the queries in the surveys described above; Rad21/SCC1 (GenBank accession number O60216), STAG1/SCC3 (Q8WVM7), CAP-H (Q15003), CAP-D2 (Q15021), CAP-G (Q9BPX3), Nse1 (Q8WV22), Nse2 (Q96MF7), Nse3 (Q96MG7), and Nse4 (Q9NXX6).

ACKNOWLEDGEMENTS

We thank Dr. T. Nakayama (University of Tsukuba, Japan) for his comments and suggestions to the manuscript. This work was supported by the grants from the Japanese Society for Promotion of Sciences awarded to Y.I. (numbers 18KK0203 and 19H03280).

REFERENCE

Anderson DE, Losada A, Erickson HP, Hirano T. 2002. Condensin and cohesin display different arm conformations with characteristic hinge angles. *J. Cell Biol.* 156(3):419-424.

Andrews EA, Palecek J, Sergeant J, Taylor E, Lehmann AR, Watts FZ. 2005. Nse2, a component of the Smc5-6 Complex, Is a SUMO ligase required for the response to DNA damage. *Mol. Cell. Biol.* 25(1):185-196.

Aragón L. 2018. The Smc5/6 complex: New and old functions of the enigmatic long-distance relative. *Annu. Rev. Genet.* 52(1):89-107.

Birkenbihl RP, Subramani S. 1995. The rad21 gene product of *Schizosaccharomyces pombe* is a nuclear, cell cycle-regulated phosphoprotein. *J. Biol. Chem.* 270(13):7703-7711.

Britton RA, Lin DC, Grossman AD. 1998. Characterization of a prokaryotic SMC protein involved in chromosome partitioning. *Genes Dev.* 12(9):1254-1259.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinform.* 10(1):421.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25(15):1972-1973.

Carramolino L, Lee B, Zaballos A, Peled A, Barthelemy I, Shav-Tal Y, Prieto I, Carmi P, Gothelf Y, González de Buitrago G, Aracil M, Márquez G, Barbero J, Zipori D. 1997. SA-1, a nuclear protein encoded by one member of a novel gene family: molecular cloning and detection in hemopoietic organs. *Gene*. 195(2):151-159.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 34(17):884-890.

Cobbe N, Heck MM. 2000. Review: SMCs in the world of chromosome biology— from prokaryotes to higher eukaryotes. *J. Struct. Biol.* 129(2-3):123-143.

Cobbe N, Heck MM. 2004. The evolution of SMC proteins: phylogenetic analysis and structural implications. *Mol. Biol. Evol.* 21(2):332-347.

Diaz M, Pecinka A. 2018. Scaffolding for repair: understanding molecular functions of the SMC5/6 complex. *Genes*. 9(1):36.

Fousteri MI, Lehmann AR. 2000. A novel SMC protein complex in *Schizosaccharomyces pombe* contains the Rad18 DNA repair protein. *EMBO J.* 19(7):1691-1702.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 28(23):3150-3152.

Fujioka Y, Kimata Y, Nomaguchi K, Watanabe K, Kohno K. 2002. Identification of a Novel Non-structural Maintenance of Chromosomes (SMC) Component of the SMC5-SMC6 Complex Involved in DNA Repair. *J. Biol. Chem.* 277(24):21585-21591.

Funayama T, Narumi I, Kikuchi M, Kitayama S, Watanabe H, Yamamoto K. 1999. Identification and disruption analysis of the recN gene in the extremely radioresistant bacterium

Deinococcus radiodurans. *Mutat. Res.* 435(2):151-161.
Gao F, Warren A, Zhang Q, Gong J, Miao M, Sun P, Xu D, Huang J, Yi Z, Song W. 2016. The All-data-based evolutionary hypothesis of ciliated protists with a revised classification of the phylum Ciliophora (Eukaryota, Alveolata). *Sci. Rep.* 6(1):24874
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29(7):644-652.
Haering CH, Farcas A, Arumugam P, Metson J, Nasmyth K. 2008. The cohesin ring concatenates sister DNA molecules. *Nature*. 454(7202):297-301.
Hirano T, Mitchison TJ. 1994. A heterodimeric coiled-coil protein required for mitotic chromosome condensation in vitro. *Cell*. 79(3):449-458.
Hirano T, Kobayashi R, Hirano M. 1997. Condensins, chromosome condensation protein complexes containing XCAP-C, XCAP-E and a xenopus homolog of the drosophila barren protein. *Cell*. 89(4):511-521.
Hu B, Liao C, Millson SH, Mollapour M, Prodromou C, Pearl LH, Piper PW, Panaretou B. 2005. Qri2/Nse4, a component of the essential Smc5/6 DNA repair complex. *Mol. Microbiol.* 55(6):1735-1750.
Ishiguro K. 2019. The cohesin complex in mammalian meiosis. *Genes Cells*. 24(1):6-30.
Janouškovec J, Paskerova GG, Miroliubova TS, Mikhailov KV, Birley T, Aleoshin VV, Simdyanov TG. 2019. Apicomplexan-like parasites are polyphyletic and widely but selectively dependent on cryptic plastid organelles. *eLife*. 8:e49662
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30(4):772-780.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*. 25(17):2286-2288.
Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AG. 2018. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*. 564(7736):410-414.
Lehmann AR, Walicka M, Griffiths DJ, Murray JM, Watts FZ, McCready S, Carr AM. 1995. The rad18 gene of *Schizosaccharomyces pombe* defines a new subgroup of the SMC superfamily involved in DNA repair. *Mol. Cell. Biol.* 15(12):7067-7080.
Losada A, Hirano M, Hirano T. 1998. Identification of xenopus SMC protein complexes required for sister chromatid cohesion. *Genes Dev.* 12(13):1986-1997.
Löwe J, Cordell SC, van den Ent F. 2001. Crystal structure of the SMC head domain: an ABC ATPase with 900 residues antiparallel coiled-coil. *J. Mol. Biol.* 306(1):25-35.
Melby TE, Ciampaglio CN, Briscoe G, Erickson HP. 1998. The Symmetrical Structure of Structural Maintenance of Chromosomes (SMC) and MukB Proteins: Long, Antiparallel Coiled Coils, Folded at a Flexible Hinge. *J. Cell Biol.* 142(6):1595-1604.
Nguyen L, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32(1):268-274.
Niki H, Jaffé A, Imamura R, Ogura T, Hiraga S. 1991. The new gene *mukB* codes for a 177 kd protein with coiled-coil domains involved in chromosome partitioning of *E. coli*. *EMBO J*. 10(1):183-193.
Palou R, Dhanaraman T, Marrakchi R, Pascariu M, Tyers M, D'Amours D. 2018. Condensin ATPase motifs contribute differentially to the maintenance of chromosome morphology and genome stability. *PLOS Biol.* 16(6):e2003980.
Pebernard S, McDonald WH, Pavlova Y, Yates JR, Boddy MN. 2004. Nse1, Nse2, and a novel subunit of the Smc5-Smc6 complex, Nse3, play a crucial role in meiosis. *Mol. Biol. Cell*. 15(11):4866-4876.

Pebernard S, Wohlschlegel J, McDonald WH, Yates 3rd JR, Boddy MN. 2006. The Nse5-Nse6 dimer mediates DNA repair roles of the Smc5-Smc6 complex. *Mol Cell Biol.* 26(5):1617-30.

Sarai C, Tanifuji G, Nakayama T, Kamikawa R, Takahashi K, Yazaki E, Matsuo E, Miyashita H, Ishida K, Iwataki M, Inagaki Y. 2020. Dinoflagellates with relic endosymbiont nuclei as models for elucidating organellogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 117(10):5364-5375.

Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2020. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 48(D1):D9-D16.

Soppa J. 2001. Prokaryotic structural maintenance of chromosomes (SMC) proteins: distribution, phylogeny, and comparison with MukBs and additional prokaryotic and eukaryotic coiled-coil proteins. *Gene*. 278(1-2):253-264.

Sutani T, Yanagida M. 1997. DNA renaturation activity of the SMC complex implicated in chromosome condensation. *Nature*. 388(6644):798-801.

Takemata N, Samson RY, Bell SD. 2019. Physical and functional compartmentalization of archaeal chromosomes. *Cell*. 179(1):165-179.

Toth A, Ciosk R, Uhlmann F, Galova M, Schleiffer A, Nasmyth K. 1999. Yeast cohesin complex requires a conserved protein, Eco1p(Ctf7), to establish cohesion between sister chromatids during DNA replication. *Genes Dev.* 13(3):320-333.

Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* 4(1):138-147.

Yazaki E, Ishikawa SA, Kume K, Kumagai A, Kamaishi T, Tanifuji G, Hashimoto T, Inagaki Y. 2017. Global Kinetoplastea phylogeny inferred from a large-scale multigene alignment including parasitic species for better understanding transitions from a free-living to a parasitic lifestyle. *Genes Genet. Syst*. 92(1):35-42.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJ. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. 541(7637):353-358.

LEGENDS FOR FIGURES

Figure 1. Inventories of the six SMC subfamilies in eukaryotes. The SMC sequences were searched for in 59 species that represent the major taxonomic assemblages in the tree of eukaryotes. The phylogenetic relationship among the major taxonomic assemblages (left) is unsettled except the node uniting TSAR, Cryptista, Haptista, and Archaeplastida, which are so-called Diaphoretickes as a whole. On the right, the presence/absence of the six SMC subfamilies is shown by symbols for each species. Blue (filled) circles represent the full-length SMC amino acid sequences retrieved from the GenBank nr database. Grey (filled) circles represent that the

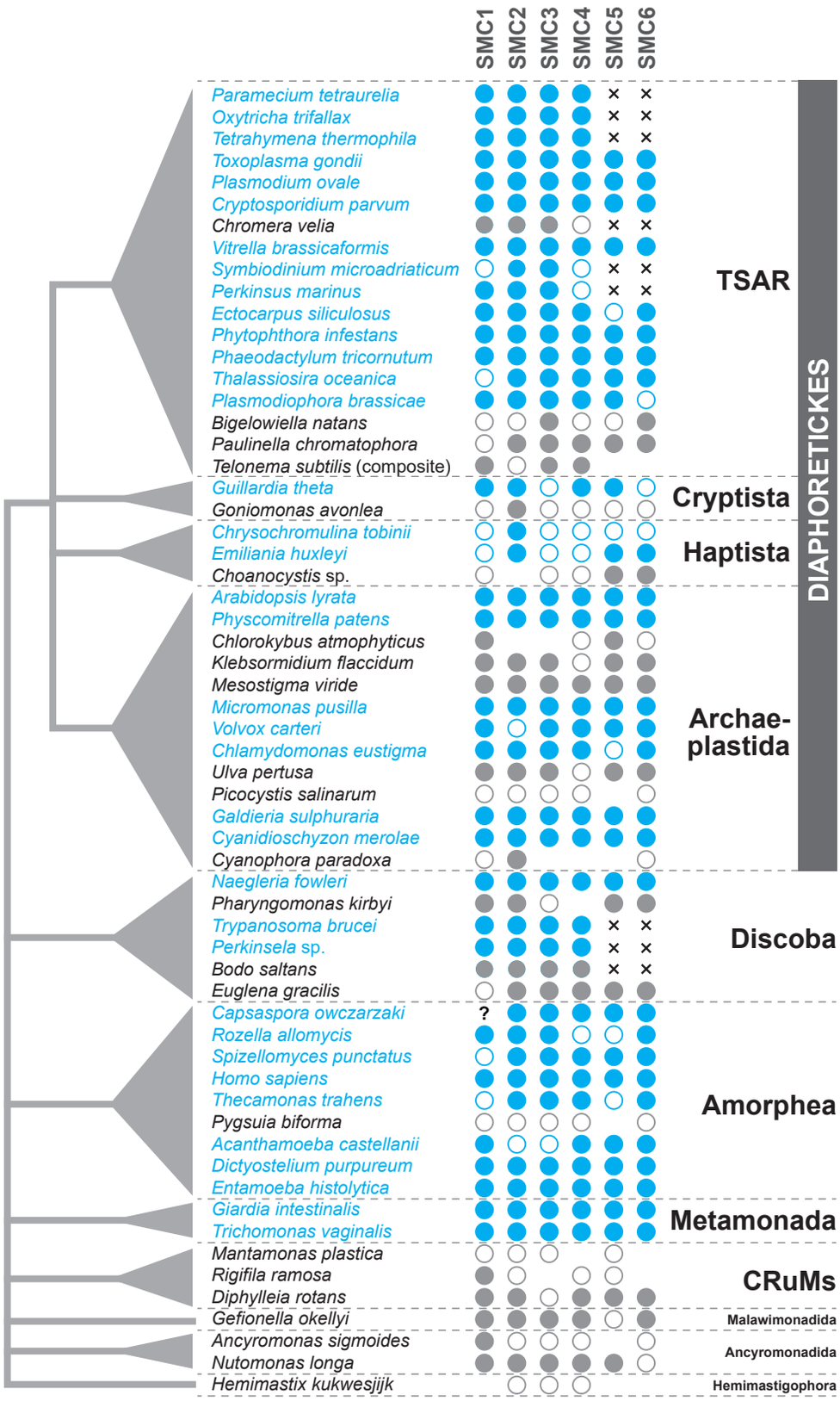
transcripts encoding the full-length proteins were found in the corresponding transcriptome data deposited in the SRA database in the GenBank. Open circles with blue borders represent truncated SMC amino acid sequences found in the GenBank nr database, while those with grey borders represent the truncated transcripts encoding SMC proteins identified in the transcriptome data. We only propose the absence of SMC5 and SMC6 in 9 eukaryotes (highlighted by x-marks), as neither of the two SMC subfamilies was found in their genome or transcriptome data. In case of a certain subfamily being undetected in the transcriptome data, we left its presence/absence uncertain and gave no symbol in the corresponding column in the figure. SMC1 in *Capsaspora owczarzaki* could be found for uncertain reasons, although both genome and transcriptome data are available (highlighted by a question mark).

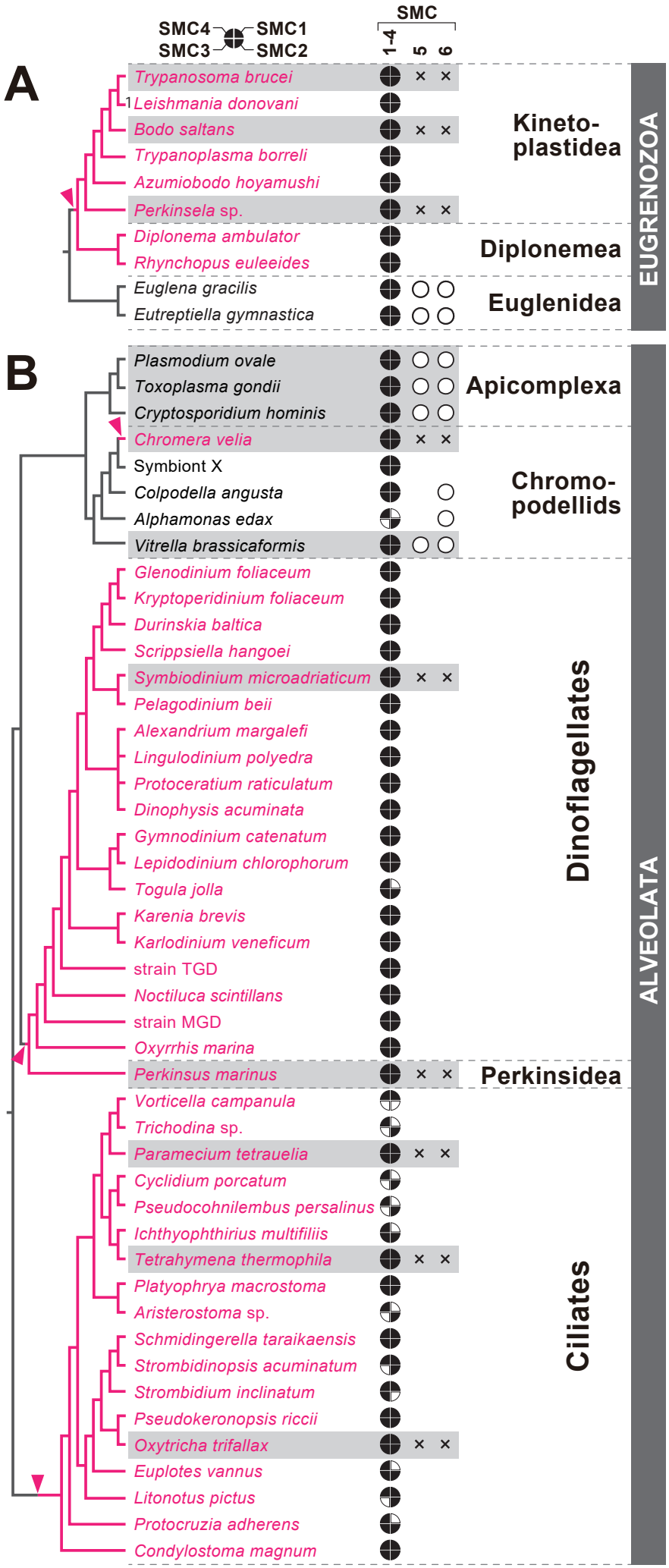
Figure 2. Taxonomic groups lacking SMC5 and SMC6. For each species examined here, the presence of SMC1-4 is represented by filled portions of the circle (upper right, SMC1; lower right, SMC2; lower left, SMC3; upper left, SMC4), regardless of whether they are full-length or truncated. If a subset of SMC1-4 was not detected in the transcriptome data, the corresponding portion(s) is left blank. Open circles indicate that we detected the full-length or truncated transcripts encoding SMC5/SMC6. According to the rule described in the legend for Figure 1, we judged the absence of SMC5 and/or SMC6 (highlighted by x-marks). In case of no transcript encoding SMC5/SMC6 being detected in a certain transcriptome, the corresponding column is left blank. The species, for which both genome and transcriptome data are available, are shaded in grey. The secondary losses of SMC5 and SMC6 occurred in the branches highlighted by arrowheads. The descendants of the organisms, which discarded SMC5 and SMC6 secondarily, are colored in magenta. (A) Inventories of SMC1-6 in the members of Euglenozoa. The phylogenetic relationship among the kinetoplastids considered here is based on Yazaki et al. (2017). (B) Inventories of SMC1-6 in Alveolata. The phylogenetic relationship among the apicomplexans and chrompodellids, that among dinoflagellates, and that among ciliates are based on Janouškovec et al. (2019), Sarai et al. (2020), and Gao et al. (2016), respectively.

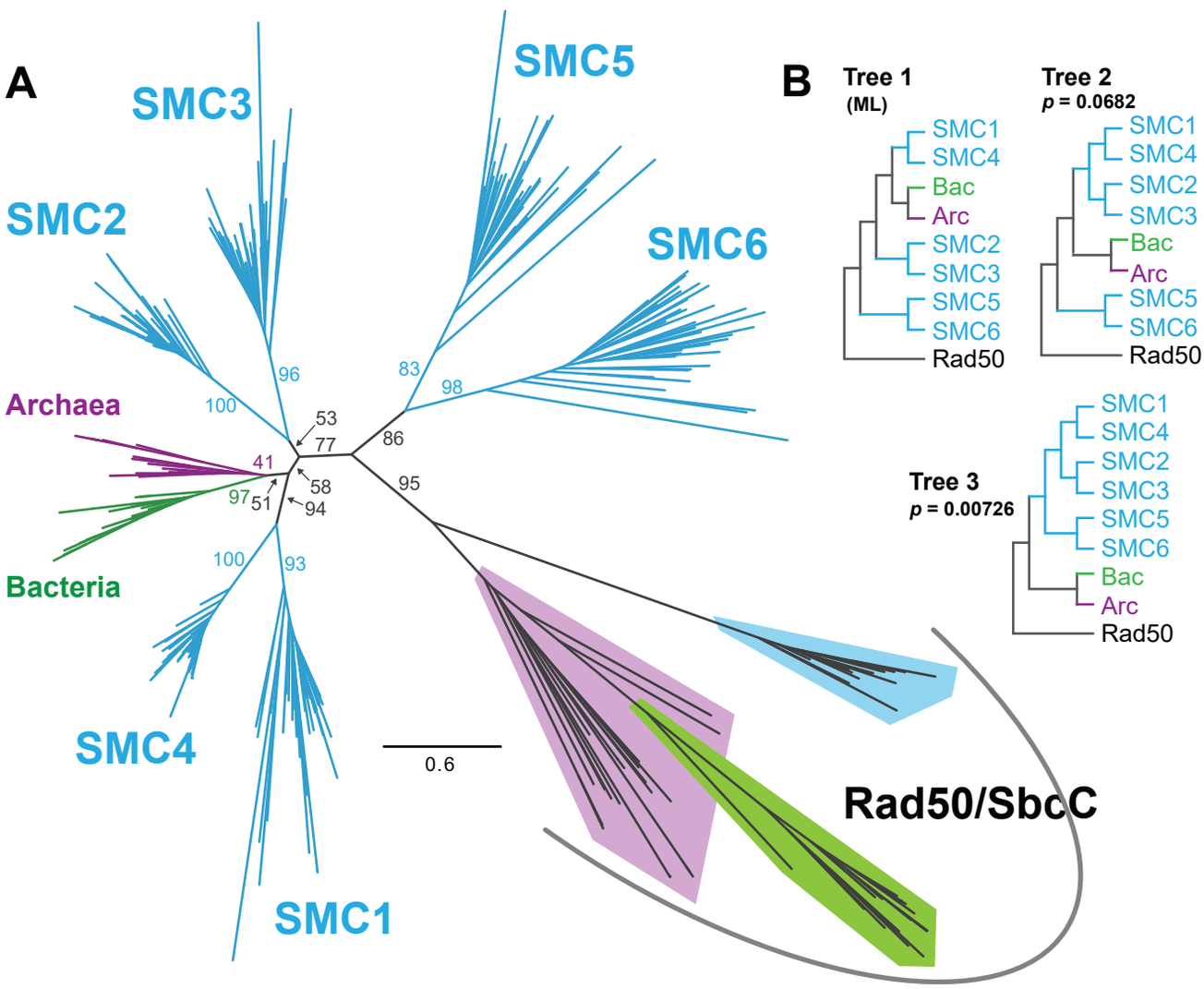
Figure 3. (A) Maximum likelihood (ML) phylogeny of SMC sequences in Bacteria, Archaea, and Eukaryota rooted by Rad50/SbcC sequences. Only ML bootstrap values for the major nodes are shown. The clade of the bacterial SMC sequences (labeled as "Bacteria") and that of the archaeal SMC sequences (labeled as "Archaea") are colored in green and purple, respectively. The clades of the six SMC subfamilies in eukaryotes (labeled as "SMC1-6") are colored in blue. In the clade of Rad50/SbcC sequences, the eukaryotic, archaeal, and bacterial branches are shaded in blue, purple, and green, respectively. (B) The results of an approximately unbiased test comparing the ML tree and two trees that represent alternative relationships among the six SMC subfamilies in eukaryotes (labeled as SMC1-6) and the "clan" containing the bacterial and archaeal SMC clades (labeled as Bac and Arc, respectively). Rad50/SbcC sequences were used as the outgroup (labeled as Rad50). The coloring scheme is the same as (A). "Tree 1" corresponds to the ML tree shown in (A). "Tree 2" was generated from Tree 1 by pruning and regrafting the "clan" containing SMC2 and SMC3 at the base of the "clan" containing SMC1 and SMC4. Tree 2 was further modified to obtain "Tree 3" in which the monophyly of SMC1-6 was enforced. The *p* values for the two alternative trees are presented.

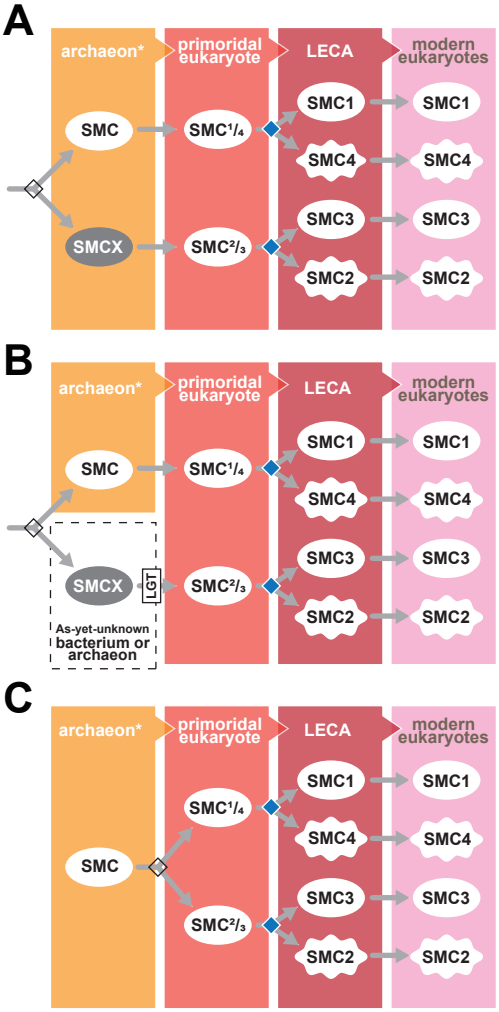
Figure 4. Scenarios for the origins and evolution of SMC1-4. We here hypothesize the evolutionary trajectories of SMC1-4 in modern eukaryotes (the rightmost column), which were descended from the last eukaryotic common ancestor (LECA) (the second rightmost column). Because of the phylogenetic affinity between SMC1 and SMC4, and that between SMC2 and SMC3 (see Fig. 3A), the ancestral molecule for SMC1 and SMC4 and that for SMC2 and SMC3 may have existed in the primordial eukaryote ("SMC¼" and "SMC⅔"; the second leftmost column). Prior to the LECA, two separate gene duplication events followed by functional divergence (blue diamonds) yielded SMC1-4. Depending on the phylogenetic relationship between SMC1-4 in eukaryotes and the bacterial/archaeal SMC, multiple scenarios for the origins of SMC¼ and SMC⅔ are possible. (A) A scenario assuming the ML tree topology in which SMC1 and SMC4 bear a specific affinity to the bacterial/archaeal SMC instead of SMC2 and SMC3 (Tree 1 in Fig. 2B). As none of the SMC and SMC-related proteins known to data shows the

direct phylogenetic affinity to SMC2 and SMC3, we need to invoke a hypothetical molecule, “SMCX.” SMCX and the authentic SMC in bacteria and archaea were derived from an ancestral SMC via gene duplication followed by functional divergence (open diamond). This particular scenario assumes that SMC¼ and SMC⅔ in the primordial eukaryote were derived vertically from the authentic SMC and SMCX in “archaeon*,” from which the first eukaryote arose (the leftmost column). (B) A scenario assuming the ML tree topology plus the lateral transfer of the gene encoding SMCX. In this scenario, SMC⅔ in the primordial eukaryote was donated from an as-yet-unknown bacterium or archaeon via lateral gene transfer (LGT), albeit SMC¼ was derived from the authentic SMC in archaeon*. (C) A scenario assuming the monophyly of SMC1-4 (Tree 2 in Fig. 3B). In this scenario, SMC¼ and SMC⅔ in the primordial eukaryote emerged from the authentic SMC in archaeon* via gene duplication followed by functional divergence (open diamond).


SMC1
SMC2
SMC3
SMC4
SMC5
SMC6
Paramecium tetraurelia
Oxytricha trifallax
Tetrahymena thermophila
Toxoplasma gondii
Plasmodium ovale
Cryptosporidium parvum
Chromera velia
Vitrella brassicaformis
Symbiodinium microadriaticum
Perkinsus marinus
Ectocarpus siliculosus
Phytophthora infestans
Phaeodactylum tricornutum
Thalassiosira oceanica
Plasmodiophora brassicae
Bigelowiella natans
Paulinella chromatophora
Telonema subtilis (composite)
Guillardia theta
Goniomonas avonlea
Chrysochromulina tobinii
Emiliania huxleyi
Choanocystis sp.
Arabidopsis lyrata
Physcomitrella patens
Chlorokybus atmophyticus
Klebsormidium flaccidum
Mesostigma viride
Micromonas pusilla
Volvox carteri
Chlamydomonas eustigma
Ulva pertusa
Picocystis salinarum
Galdieria sulphuraria
Cyanidioschyzon merolae
Cyanophora paradoxa
Naegleria fowleri
Pharyngomonas kirbyi
Trypanosoma brucei
Perkinsela sp.
Bodo saltans
Euglena gracilis
Capsaspora owczarzaki
Rozella allomycis
Spizellomyces punctatus
Homo sapiens
Thecamonas trahens
Pygsuia biforma
Acanthamoeba castellanii
Dictyostelium purpureum
Entamoeba histolytica
Giardia intestinalis
Trichomonas vaginalis
Mantamonas plastica
Rigifila ramosa
Diphylleia rotans
Gefionella okellyi
Ancyromonas sigmoides
Nutomonas longa
Hemimastix kukwesjijk
TSAR
Cryptista
Haptista
Archae-
plastida
Discoba
Amorphea
Metamonada
CRuMs
Malawimonadida
Ancyromonadida
Hemimastigophora
DIAPHORETICKES


SMC4
SMC3
SMC1
SMC2
SMC
1-4
5
6
A
Trypanosoma brucei
Leishmania donovani
Bodo saltans
Trypanoplasma borreli
Azumiobodo hoyamushi
Perkinsela sp.
Diplonema ambulator
Rhynchopus euleeides
Euglena gracilis
Eutreptiella gymnastica
Kineto-plastidea
Diplonemea
Euglenidea
EUGRENOZOA
B
Plasmodium ovale
Toxoplasma gondii
Cryptosporidium hominis
Apicomplexa
Chromera velia
Symbiont X
Colpodella angusta
Alphamonas edax
Vitrella brassicaformis
Chromo-podellids
Glenodinium foliaceum
Kryptoperidinium foliaceum
Durinskia baltica
Scrippsiella hangoei
Symbiodinium microadriaticum
Pelagodinium beii
Alexandrium margalefi
Lingulodinium polyedra
Protoceratium raticulatum
Dinophysis acuminata
Gymnodinium catenatum
Lepidodinium chlorophorum
Togula jolla
Karenia brevis
Karlodinium veneficum
strain TGD
Noctiluca scintillans
strain MGD
Oxyrrhis marina
Dinoflagellates
Perkinsus marinus
Perkinsidea
Vorticella campanula
Trichodina sp.
Paramecium tetrauelia
Cyclidium porcatum
Pseudocohnilembus persalinus
Ichthyophthirius multifiliis
Tetrahymena thermophila
Platyophrya macrostoma
Aristerostoma sp.
Schmidingerella taraikaensis
Strombidinopsis acuminatum
Strombidium inclinatum
Pseudokeronopsis riccii
Oxytricha trifallax
Euplotes vannus
Litonotus pictus
Protocruzia adherens
Condylostoma magnum
Ciliates
ALVEOLATA


A
SMC3
SMC5
SMC2
SMC6
Archaea
Bacteria
SMC4
SMC1
Rad50/SbcC
0.6
96
100
83
98
53
77
86
41
58
95
97
51
94
100
93
B
Tree 1
(ML)
SMC1
SMC4
Bac
Arc
SMC2
SMC3
SMC5
SMC6
Rad50
Tree 2
p = 0.0682
SMC1
SMC4
SMC2
SMC3
Bac
Arc
SMC5
SMC6
Rad50
Tree 3
p = 0.00726
SMC1
SMC4
SMC2
SMC3
SMC5
SMC6
Bac
Arc
Rad50


A
archaeon*
primoridal eukaryote
LECA
modern eukaryotes
SMC
SMCX
SMC1/4
SMC2/3
SMC1
SMC4
SMC3
SMC2
SMC1
SMC4
SMC3
SMC2
B
archaeon*
primoridal eukaryote
LECA
modern eukaryotes
SMC
SMCX
LGT
As-yet-unknown bacterium or archaeon
SMC1/4
SMC2/3
SMC1
SMC4
SMC3
SMC2
SMC1
SMC4
SMC3
SMC2
C
archaeon*
primoridal eukaryote
LECA
modern eukaryotes
SMC
SMC1/4
SMC2/3
SMC1
SMC4
SMC3
SMC2
SMC1
SMC4
SMC3
SMC2