Shaping up genomes: Prokaryotic roots and eukaryotic diversification of SMC complexes

Jolien J.E. van Hooff[1,2,7]*, Maximilian W.D. Raas[3,4,5], Eelco C. Tromer[6] and Laura Eme[1]*

1. Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech Gif-sur-Yvette, France
2. Laboratory of Microbiology, Wageningen University and Research, The Netherlands
3. Theoretical Biology and Bioinformatics, Faculty of Science, Utrecht University, The Netherlands
4. Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW), The Netherlands
5. Oncode Institute, Utrecht, The Netherlands
6. Cell Biochemistry, Groningen Biomolecular Sciences & Biotechnology Institute, Faculty of Science and Engineering, University of Groningen, The Netherlands

***Correspondence to:** jolien.vanhooff@wur.nl, laura.eme@universite-paris-saclay.fr

Key words: genome organization, Structural Maintenance of Chromosomes, LECA, condensin, cohesin, SMC5/6, evolutionary cell biology, eukaryogenesis, chromatin

Summary

Across the tree of life, SMC complexes organize, segregate and regulate DNA. In contrast to prokaryotes, which often possess one SMC complex, eukaryotes usually have four such complexes. This expanded set is involved in managing the considerably larger nuclear genomes of eukaryotes, which are distributed across multiple chromosomes. Despite their essential functions, SMC complexes exhibit variations across model eukaryotes, suggesting an even greater variety of these complexes that remains unexplored. Here, we aimed to uncover the diversity and evolution of SMC complexes across eukaryotes, and their deeper, prokaryotic evolutionary origins. For this, we conducted in-depth comparative genomic and phylogenetic analyses of SMC complexes. We show that the last eukaryotic common ancestor (LECA) likely had fully-fledged versions of all four complexes and that the condensin II complex was later lost at least 30 times in eukaryotes. We report evidence that proteins previously designated as functional analogs in various model organisms (e.g., Sororin, Securin, Nse5 and Nse6) are in fact genuine orthologs. Finally, we traced the prokaryotic origins of these complexes and propose that a single SMC complex duplicated in an early archaeon. Altogether, we provide a comprehensive overview of eukaryotic SMC complex diversity and evolution, both addressing and generating questions about their functioning in ancestral and contemporary organisms.

Introduction

The organization of large eukaryotic nuclear genomes is a complex and dynamic process[1]. At the highest level, a typical mammalian genome folds into chromosome territories within the nucleus. Chromosomes themselves consist of two main types of regions: transcriptionally active euchromatin and silenced heterochromatin. These regions tend to cluster with their equivalents into A and B compartments, respectively. At a smaller scale, DNA forms topologically associating domains (TADs), which regulate the expression of some genes[2], although their exact activity remains unclear. At the lowest level, DNA is wrapped around a histone octamer, forming a nucleosome.

SMC (Structural Maintenance of Chromosome) complexes play a crucial role in genome organization and stability across all domains of life[3]. Eukaryotic and prokaryotic SMC complexes share a common basic structure, with two SMC proteins that together form a composite ATPase, and dimerize via their hinge domains (Figure 1). A long and flexible protein called kleisin connects their heads by binding asymmetrically to one SMC protein's neck, and to the other's cap. The resulting tripartite ring is supplemented with additional proteins: HEAT(Huntingtin-EF3-PP2A-TOR1)-repeat-containing hawks in the eukaryotic condensin and cohesin, or winged-helix domain-containing kites in prokaryotic SMC complexes and the eukaryotic SMC5/6 complex. Eukaryotic SMC complexes also incorporate a variable number of accessory, sometimes transient subunits, which regulate their activity and localization. Using ATP, SMC complexes can catch and move DNA through their structures to build and progressively enlarge DNA loops, a process referred to as loop extrusion[2].

Eukaryotes typically possess four SMC complexes: condensin I, condensin II, cohesin, and SMC5/6 (Figure 1), whereas many prokaryotes have just one. The eukaryotic SMC complexes probably emerged during eukaryogenesis, the evolutionary transition from prokaryotic to eukaryotic cells. It remains unclear which evolutionary events underlie the origination of the eukaryotic SMC complexes and how this origination might be linked to the emergence of eukaryotic chromosome features. Previous work suggested that two different SMC complexes were inherited from the Asgard archaeal-related eukaryotic ancestor[4–6], giving rise to condensin plus cohesin, and to SMC5/6, respectively[7]. Furthermore, not all eukaryotes have retained the SMC complexes. For example, several lineages lost SMC5/6 and condensin II[7–9], whereas ciliates expanded condensin I[10]. Interestingly, some of these evolutionary dynamics seem linked to genome organization, as species lacking condensin II appear to lack chromosome territories[9], although the global extent of this pattern remains to be tested.

Here, our study sought to thoroughly examine eukaryotic SMC complexes, aiming to uncover their evolution both before and after the last eukaryotic common ancestor (LECA) using state-of-the-art homology detection and molecular phylogenetics on a broad taxon sampling. Specifically, we aimed to elucidate how SMC complexes diversified after LECA, reconstruct LECA's SMC complexes, and provide insights into how LECA acquired these complexes during eukaryogenesis. To achieve this, we conducted a comprehensive analysis of the proteomes of 177 diverse eukaryotes and 4,101 prokaryotes, including 28 Asgard archaea, the closest living prokaryotic relatives of eukaryotes[6]. Based on these analyses, we present a new case study on how eukaryotic protein complexes emerged and evolved, offer hypotheses about genome organization in non-model eukaryotes, and provide a valuable resource for in-depth functional investigation of individual proteins based on their sequence evolution.

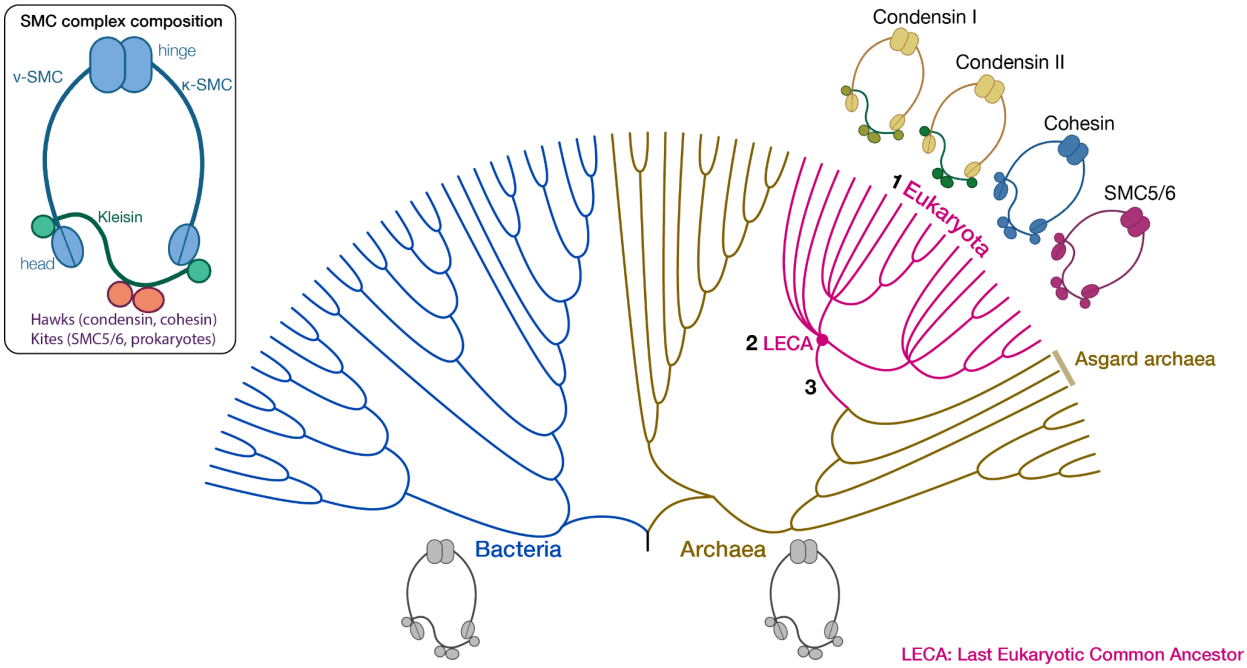


Figure 1. The tree of life and SMC complexes in the three major groups: Bacteria, Archaea and Eukaryota. Eukaryotes typically have four SMC complexes: condensin I, condensin II, cohesin and

SMC5/6. This study addresses three questions: (1) What is the diversity of SMC complexes among eukaryotes? (2) What were the characteristics of the SMC complexes in the Last Eukaryotic Common Ancestor (LECA)? (3) How did SMC complex subunits originate before LECA? Inset: general composition and architecture of SMC complexes across the tree of life. The SMC proteins form dimers *via* their hinge domains (top) and their ATPase heads are linked to the kleisin protein (bottom). Between the head and hinge domains, they feature a backfolded coiled-coil region. The kleisin attaches to one SMC protein through its N-terminus, binding the SMC protein at its neck (the ν-SMC), and to the other SMC protein via its C-terminus, binding it at the SMC protein's cap (the κ-SMC). Note that in most prokaryotic SMC complexes, the ν-SMC and κ-SMC are the same protein, which form a homodimer. In eukaryotic SMC complexes, these are different proteins forming a heterodimer. In addition to the SMC and kleisin proteins, SMC complexes generally include two core other subunits, either kite proteins (in prokaryotic SMC complexes and the eukaryotic SMC5/6 complex) or hawk proteins (in the eukaryotic condensin and cohesin complexes). While not depicted here, in addition to the core subunits, eukaryotic SMC complexes often have accessory members, or more transient interactors. Created with BioRender.com.

Results

Condensin II was lost recurrently in eukaryotic evolution

The condensin complexes condense chromosomes during mitosis[11]. However, the influence of condensin II extends to interphase, where it appears to induce the formation of a nuclear genome architecture known as “chromosome territories”. This architecture involves chromosomes occupying distinct regions within the nucleus[12]. This effect seems consistent across species, as those lacking condensin II exhibit a different nuclear arrangement known as a ‘Rabl-like’ architecture, with the centromeres and/or telomeres of different chromosomes clustering together[9].

Condensin I and II share common SMC subunits (SMC2 and SMC4) but differ in their kleisin proteins (CAPH *versus* CAPH2) and hawk proteins (CAPG and CAPD2 *versus* CAPG2 and CAPD3). We searched for orthologs of condensin I and II subunits across a diversity of eukaryotic representatives. We found that SMC2 and SMC4 are widely conserved, suggesting that species possess at least one condensin complex (Figure 2, Supplementary Figure 1, Supplementary Table 1). Our analysis also revealed that the subunits specific to each complex, i.e. CAPH, CAPG, and CAPD2 (condensin I), and CAPH2, GAPG2, and CAPD3 (condensin II), are coherently present or absent in most species, indicating that they co-evolve. A notable exception to this involves the condensin II complex in insects, where only one or two subunits were repeatedly lost[8].

By mapping the absences of complex-specific subunits, we inferred at least six losses of condensin I and a substantial 30 losses of condensin II. The absence of individual proteins should be interpreted carefully, since they could be due to extensive sequence divergence or incomplete protein datasets. Because of this uncertainty, we only considered a condensin complex absent if, in a given lineage, all three complex-specific proteins were undetected. Contrary to previous assumptions[11], we observed that certain lineages exclusively possess

condensin II, as exemplified by the human pathogen *Trichomonas vaginalis*. The remarkably high frequency of condensin II losses prompts an investigation into whether these losses consistently correlate with transitions from chromosome territories to Rabl-like interphase nuclear architectures[9]. Of note, whereas a lack of condensin II seems coupled to a Rabl-like architecture, the reverse may not hold across lineages. For example, malaria-causing *Plasmodium falciparum* encodes condensin II, yet exhibits such a Rabl-like architecture[13,14] (Supplementary Figure 1, Supplementary Table 1).

Additionally, while condensin II was thought to be absent in fungi[15], we detected its presence in the fungus *Gonapodya prolifera*. Considering that condensin II is also found in fungal relatives like the nucleariid *Parvularia atlantis*, it may still have been present in the last common ancestor of Fungi, and subsequently lost in multiple fungal lineages. An alternative scenario is that condensin II was lost before the last common ancestor of Fungi and later acquired by *Gonapodya* through horizontal gene transfer (HGT) of all subunits. However, this HGT event does not seem supported by our gene phylogenies (Supplementary Data). To validate the unexpected presence of condensin II in *G. prolifera*, we examined its sister species *Gonapodya sp.* JEL0774, in which we also detected all three subunits. This confirms that the detection of this fungal condensin II is not due to contamination. *Gonapodya* species' possession of condensin II raises the question of whether they adopt a chromosome territories-like architecture.

We also observed instances of condensin expansions, such as in ciliates, which expanded their condensin I repertoire in the absence of condensin II, confirming observations in the ciliate *Tetrahymena thermophila*[10,16]. Although the phylogenies of CAPH, CAPG and CAPD2 proteins do not point to one specific scenario (Supplementary Data), they suggest that these proteins may have duplicated in parallel across different ciliate lineages. Conceivably, after condensin II was lost in the ciliate ancestor (Supplementary Figure 1), multiple lineages independently expanded their condensin I repertoire, which functionally differentiated, as observed in *T. thermophila*[10]. These expanded repertoires may enable ciliates to independently orchestrate their two distinct genomes - somatic and germline.

The canonical roles of condensin I and condensin II, and to what extent these overlap, remain unresolved issues. In vertebrate model organisms, condensin I is known to operate during mitosis by laterally compacting chromosomes and reducing their width, while condensin II starts its action already in interphase, axially shortening chromosomes[11,17]. Yet, this functional differentiation does not appear consistent across species[11]. For example, the red alga *Cyanidioschyzon merolae* exhibits dynamics of condensin I and II that are distinct from what is known in mammals[18]. Given the presence of both complexes across distantly related eukaryotes (Figure 2), an experimental comparative approach would be powerful in discerning their various roles within eukaryotic cells.

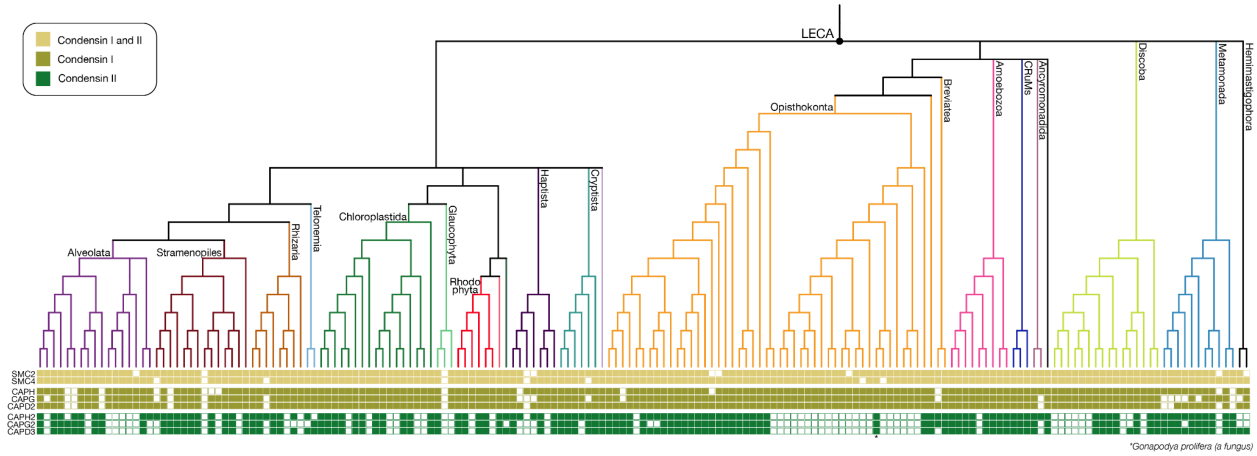


Figure 2. Distributions of condensin I and II subunits across the eukaryotic tree of life. SMC proteins: SMC2,4. Kleisin proteins: CAPH (condensin I), CAPH2 (condensin II). Hawk proteins: CAPG and CAPD2 (condensin I), CAPG2 and CAPD3 (condensin II).

LECA harbored a complete SMC5/6 complex, including Nse5 and Nse6

The SMC5/6 complex stands out as the most cryptic eukaryotic SMC complex. It is thought to be implicated in a variety of cellular processes, like genome replication, DNA damage repair, and defense against viral DNA[19]. Unlike the condensin and cohesin complexes, which contain hawk proteins, the SMC5/6 complex contains kite proteins; hawks and kites, however, do not appear to be homologous[20].

A recent study concluded that the SMC5/6 complex was lost at least four times across eukaryotes[7]. Our analysis, utilizing a more comprehensive dataset, indicates that the complex has been lost at least eight times, even if we strictly consider the absence of all core subunits (SMC5, SMC6, kleisin Nse4, and kites Nse1 and Nse3) as the loss of the complex (Figure 3, Supplementary Figure 1, Supplementary Table 1). Most of these losses seem ancient, having taken place in the deeper branches of the eukaryotic tree of life, such as in the ancestor of ciliates. Yet, some losses occurred more recently, such as in the metazoan blood fluke *Schistosoma mansoni* and tardigrade *Ramazzottius varieornatus.*

We further noticed that the primary subunits, and the associated SUMO ligase Nse2, co-occur across species, suggesting that they co-evolve. From an evolutionary perspective, Nse2 thus also belongs to the SMC5/6 complex core.

Nse5 and Nse6 are often regarded as taxonomically restricted and divergent components of the SMC5/6 complex, with Nse5 orthologs being particularly challenging to detect[21]. Direct sequence comparisons fail to reveal homology between the experimentally studied yeast protein Nse5 and its human analog, SLF1. However, employing sensitive sequence searches, and structure prediction and similarity searches, we established that Nse5 analogs in various model organisms are homologous, and the same applies to Nse6. Remarkably, our analyses have also

identified a putative Nse5 in *Drosophila melanogaster*—a species previously thought to lack both Nse5 and Nse6[22].

Overall, we detected Nse5 and Nse6 across eukaryotic diversity, pointing to their presence in LECA, yet only in modern species possessing other SMC5/6 subunits. The distribution of Nse5 is remarkably sparse, but this probably in part reflects our inability to detect these highly divergent orthologs. Notably, Nse5 orthologs diverged via the acquisition of additional domains, such as ankyrin and BRCT repeats[23]. Moreover, Nse5 and Nse6 adopt a similar alpha-solenoid structure (Supplementary Figure 4), which probably indicates their shared evolutionary origin. We propose they resulted from a single Nse5,6 ancestor protein that duplicated before LECA (Supplementary Text).

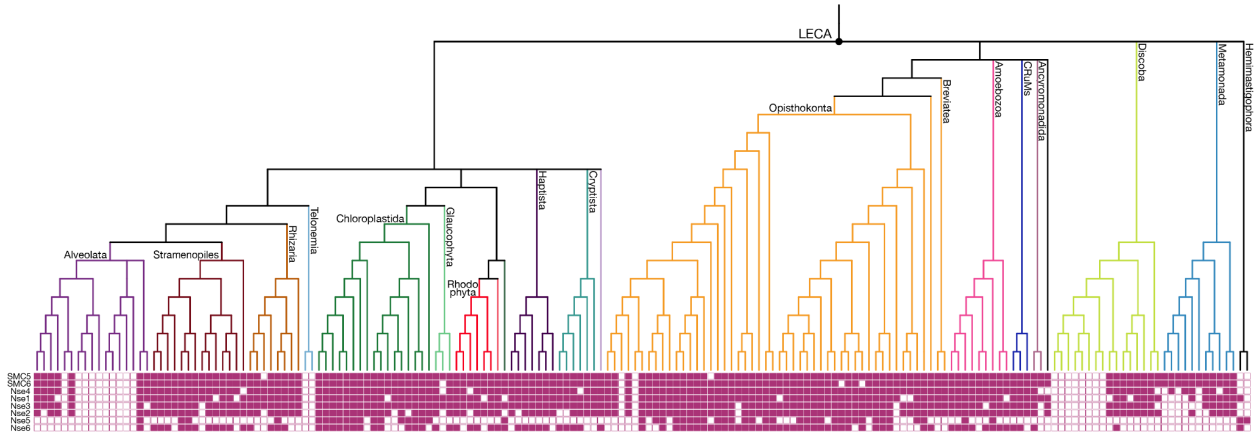


Figure 3. Distributions of SMC5/6 complex subunits across the eukaryotic tree of life. SMC proteins: SMC5,6. Kleisin protein: Nse4. Kite proteins: Nse1,3. Other subunits: Nse2,5,6.

Cohesin displays a mosaic of divergent members and partners

The cohesin complex fulfills two critical roles. First, it maintains sister chromatids together from DNA replication until anaphase onset, ensuring chromosomes are properly aligned and attached to the mitotic spindle for safe segregation. Second, at least in vertebrates, cohesin organizes DNA into loops, forming TADs. Within these TADs, sequences interact more frequently with each other than with others, potentially influencing gene regulation[2]. Cohesin's composition is more intricate than that of condensin, characterized by a diverse array of both core and peripheral subunits. The presence of SMC1 and SMC3 across eukaryotes suggests a near-ubiquitous distribution of cohesin (Figure 4, Supplementary Figure 1, Supplementary Table 1). However, the distribution of non-SMC proteins is more sporadic and inconsistent, with different subunits absent in various lineages, obscuring any clear functional modules among these components.

Cohesin includes two kleisins, Scc1 and Rec8, functioning in mitosis and meiosis, respectively[24]. The wide but patchy distributions of these paralogs, combined with their phylogeny (Figure 4), suggest that Scc1 and Rec8 resulted from a pre-LECA gene duplication, that both were present in LECA, and were subsequently lost multiple times independently. The limited sequence divergence of Scc1 and Rec8 suggests their functions might have remained

closely related, with the mitosis-meiosis differentiation possibly incomplete in LECA. This in turn could explain the frequent loss of one of the two paralogs, due to their potential functional redundancy.

The cohesin complex also contains three hawk proteins: Scc3, PDS5, and NIPBL. PDS5 and NIPBL appear to exclude each other from binding to the kleisin[25,26]. All three hawks seem relatively sparsely distributed, raising the hypothesis that they are partially functionally redundant. Yet, various lineages have multiple paralogs of individual hawks, such as humans, which have paralogs of Scc3 (SA1/2/3) and PDS5 (PDS5A/B). These paralogs were derived from recent gene duplications, and we infer they descend from a single protein in LECA.

Finally, cohesin has various modulators, like Securin and Sororin, both crucial for sister chromatid cohesion. Our reevaluation of Securin's characterized analogs in different model organisms confirmed their homology[27], and Securin's widespread occurrence. Similarly, we confirm the homology of Sororin candidates[28], and show its broad distribution, despite them being largely disordered with highly divergent sequences. We identified Securin and Sororin orthologs mostly based on short linear motifs, which are not always unique, complicating homology detection and raising the possibility of convergent evolution in some cases (Supplementary Text).

The broad presence of Securin and Sororin suggests that they already regulated LECA's cohesin complex. In contrast, CTCF, another cohesin regulator, appears to be animal-specific while being part of a large zinc finger protein family (Supplementary Data). CTCF exemplifies a common evolutionary strategy in eukaryotic protein complexes: conserving a core structure while acquiring novel components in specific lineages[29].

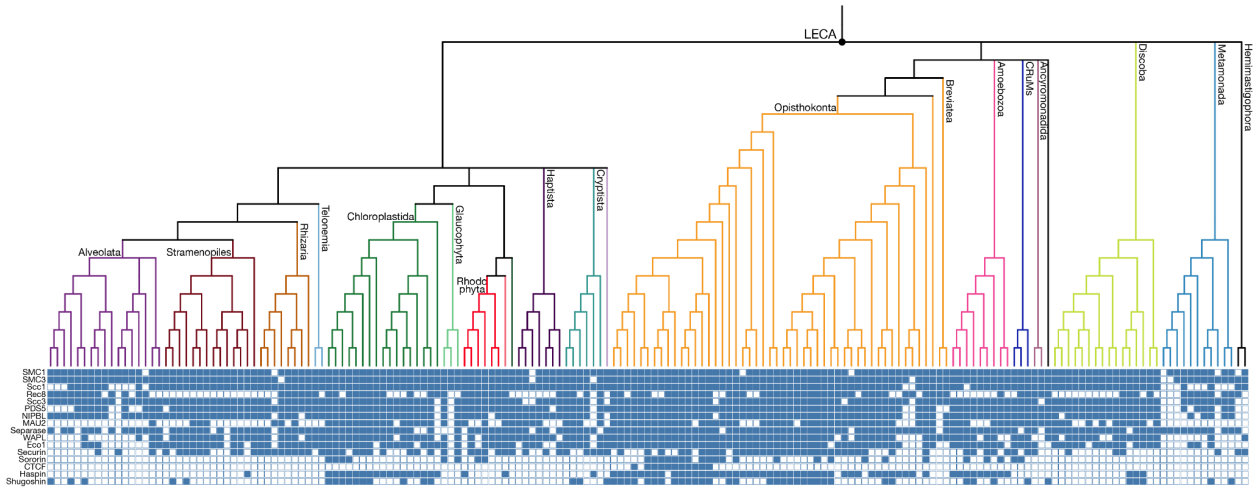


Figure 4. Distributions of cohesin subunits across the eukaryotic tree of life. SMC proteins: SMC1,3. Kleisin proteins: Scc1, Rec8. Hawk proteins: Scc3, PDS5, NIPBL. Other (transient) subunits and interactors: MAU2, WAPL, Eco1, Securin, Sororin, Haspin, Shugoshin, Separase, CTCF.

Eukaryotes inherited an expanded SMC array from their archaeal ancestors

To understand the origin of the four SMC complexes during eukaryogenesis, we delved into the evolutionary history of their subunits. Our goal was to clarify their evolutionary relationships among each other and to trace their prokaryotic roots. Building upon Yoshinaga & Inagaki's hypothesis that condensin and cohesin originated from a different prokaryotic SMC complex than SMC5/6[7], we expanded the investigation by incorporating a broader range of species and by (re)examining the non-SMC proteins of the kleisin and kite families. Phylogenetic analysis of hawks was not feasible due to their extensive divergence[20], but we nonetheless investigated their origins using complementary non-phylogenetic approaches.

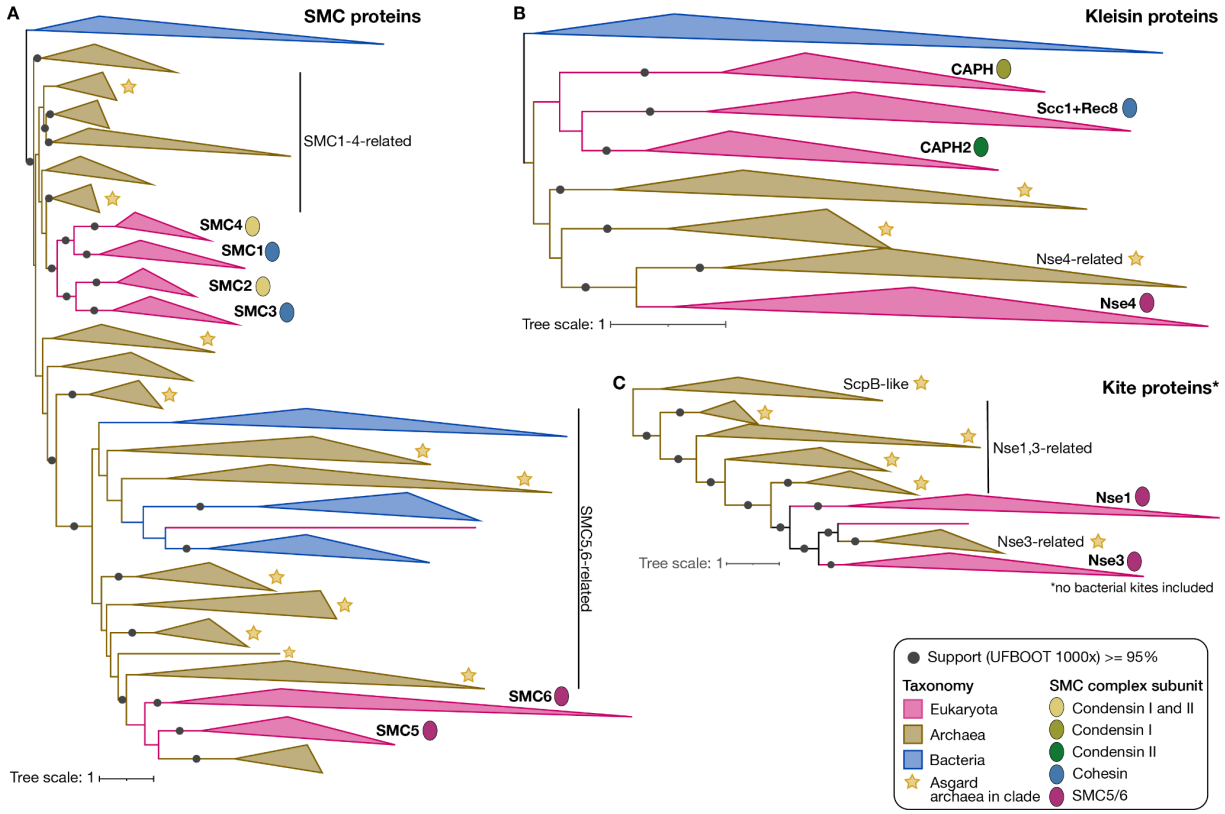


Figure 5. A-C. Phylogenetic trees of SMC (A), kleisin (B) and kite (C) proteins inferred using IQ-TREE[30] under the LG+C60+G mixture model. Statistical support at branches was estimated from 1000 ultrafast bootstrap replicates[31]. In the kleisin tree, four sequences with very long branches (>5 average substitutions per site) and unclear affiliations were removed. The uncollapsed, unfiltered trees can be found on https://itol.embl.de/tree/1372242521226653169746804 0 (SMC), https://itol.embl.de/tree/1372242521264616974652 03 (kleisin), and https://itol.embl.de/tree/1372242521219871169752525 5 (kite). Scale bars indicate estimated average numbers of substitution per site.

SMC proteins

Our phylogenetic analysis aligns with previous findings[7] and shows that eukaryotic SMC proteins form two clusters, one constituted by condensin and cohesin SMC proteins, and one by SMC5 and SMC6, separated by several groups of prokaryotic homologs (Figure 5A). However, topology tests suggest that we cannot reject the possibility that these two clusters actually form a monophyletic group (p-AU=0.317). Both the condensin/cohesin SMC proteins and SMC5,6 show diverse archaeal homologs as their closest relatives, including Asgard archaea, indicating that eukaryotes inherited them from their archaeal ancestors (Supplementary Table 3). The SMC5,6 cluster is interspersed with a small (3 sequences) cluster of TACK archaeal sequences; given the small number, narrow distribution and low support we think this likely reflects an artifact.

Many archaea have two SMC proteins (Supplementary Table 3), likely resulting from a gene duplication event in an early archaeal ancestor. We confirm that after eukaryotes diverged from their archaeal ancestors, the ancestral condensin/cohesin SMC protein underwent an additional duplication event, resulting in the ancestors of SMC1,4 and SMC2,3. Similarly, the ancestor of SMC5,6 duplicated to give rise to SMC5 and SMC6[7]. In both cases, such a duplication resulted in a κ-SMC (Figure 1, SMC1,4, and 5) and a ν-SMC (SMC2,3, and 6). Thus, these subfunctionalizations within the SMC protein complexes occurred in parallel, yielding heterodimers of SMC proteins rather than homodimers. After this heterodimer-forming duplication, the condensin/cohesin SMC proteins duplicated once more, giving rise to the condensin and cohesin SMC heterodimers.

Kleisin proteins

In contrast with previous work[7], our phylogenies suggest that the kleisins from condensin/cohesin and from SMC5/6 do not form a monophyletic group (Figure 5B). Their monophyly was rejected by the Approximately Unbiased (AU) topology test ((p-AU=0.002). However, interpreting the phylogeny in terms of prokaryotic ancestry poses challenges, since the eukaryotic condensin/cohesin kleisins emerge as basal to all archaeal sequences with weak statistical support. However, numerous archaea from diverse superphyla–Asgard, TACK, and Euryarchaeota–are repeatedly represented in the tree (Supplementary Table 3). One of those archaeal clades represents the well-supported sister-group of the SMC5/6 kleisin Nse4, suggesting its archaeal ancestry.

In contrast, the rest of the phylogeny is poorly supported. We hypothesize that the second archaeal clade, which contains the same organismal diversity as the Nse4-related archaeal clade, represents the true sister-group of the condensin/cohesin kleisins. We tested this hypothesis by inspecting the genomic neighborhood of this group of archaeal, hypothetically condensin/cohesin-related kleisins, and we frequently found them next to the condensin/cohesin-related SMC proteins (Supplementary Text, Supplementary Figure 3). This genomic localization thus provides support for their sisterhood to eukaryotic condensin/cohesin-related kleisins. The duplication of kleisins in an archaeal ancestor would be coherent with the previously discussed duplication scenario for the SMC proteins. Within the

condensin/cohesin kleisins clade, it is not clear which kleisins are most closely related to one another, since the deeper branches are poorly supported. Given the putative close relationship of condensin I and II - they even share their SMC proteins - we consider it most plausible that CAPH and CAPH2 are most closely related to one another; a hypothesis also hinted at by a more restricted phylogenetic analysis (Methods, Supplementary Figure 2).

In summary, both eukaryotic kleisins and SMC proteins appear to have originated from an ancient archaeal gene duplication, which was followed by further duplications specific to eukaryotic evolution.

Kite proteins

The SMC5/6 complex contains kite proteins, instead of the condensin/cohesin-associated hawks[20]. Our search for kite homologs indicated three distinct sequence 'archetypes': Nse1/3-like (encompassing archaeal kites), ScpB-like (including both archaeal and bacterial kites, such as ScpB in *Bacillus subtilis*), and MukE-like (covering other bacterial kites, like MukE in *Escherichia coli*). Despite their marked sequence divergence, these archetypes share similar structural features. Our kite phylogeny mirrors the ones of SMC and kleisin proteins, where various archaea possess two copies of kite, whose placements suggest at least one duplication event in an archaeal ancestor (Figure 5C). One cluster contains a wide range of archaea from all superphyla and form a monophyletic group with eukaryotic Nse1,3, whereas the other is of the ScpB-like archetype (Supplementary Table 3). Some Asgard archaea even possess three kite paralogs, with one of them forming an exclusively Asgard archaeal clade clustering specifically with the eukaryotic Nse3. This suggests that the duplication event leading to Nse1 and Nse3 occurred in an Asgard archaeal ancestor. These paralogous Asgard kite genes co-localize in the genome with different SMC and kleisin genes, suggesting they participate in distinct SMC complexes (Supplementary Text), contrasting Nse1 and Nse3, which both participate in SMC5/6. In conclusion, archaeal kite protein phylogenies and genomic context suggest an ancient duplication event, coupled with a more recent duplication specific to a subset of Asgard archaea.

Hawk proteins

Condensin and cohesin are joined by hawk proteins, which are HEAT repeat proteins. The degenerate nature of these repeats renders their sequences too divergent for conventional phylogenetic approaches. Wells et al. developed a sequence profile network approach to dissect ancient homologous links for such repeat proteins[20]. They proposed that hawk proteins Scc3, CAPG,G2, CAPD2,3 likely form one ancestral eukaryotic family within the larger HEAT repeat family, and lack a clear prokaryotic affiliation. Using a similar approach with deeper evolutionary coverage (Methods), combined with comparisons of AlphaFold2-predicted 3D structures, we aimed to further delineate the evolutionary history of hawks, as well as other accessory repeat subunits (Supplementary Text).

Using the network approach we corroborated that the hawk proteins are paralogous to one another, and resulted from gene duplications before LECA. Evolutionarily, they form two groups, which correspond to the previously described functional groups in the condensin/cohesin

complexes[32]: the A-type hawks CAPG, CAPG2, Scc3, and the B-type hawks CAPD2, CAPD3, NIPBL and PDS5 (Figure 6A, Supplementary Tables 5,6). Indeed, structural comparisons show the same A/B-type subdivision, with the marked divergent members for each respective group being CAPG2 and PDS5, while the closest association was found between CAPD2 and CAPD3 (DALI Z-score:19.15, Figure 6B, Supplementary Table 7). This A/B division suggests that in the ancestor of condensin and cohesin, the hawk protein replaced kite(s), first forming a homodimer, and then duplicated, resulting in an A- and B-type-containing heterodimer.

Within the A- and B-type hawks, the relationships are not very clear, probably due to the extensive divergence of CAPG2 and PDS5. However, we propose this heterodimer to subsequently have duplicated to give rise to a separate condensin and cohesin dimer, because this would be most congruent with the closer relationship of condensin I and condensin II to one another. Thereafter, in condensin another duplication of the heterodimer occurred, leading to the heterodimers present in extant condensin I and II. In cohesin, also an additional duplication occurred, but only of the B-type hawk, yielding PDS5 and NIPBL. This duplication effectively yielded two cohesin hawk heterodimers, namely Scc3-PDS5 and Scc3-NIPBL, and an individual cohesin complex has been shown to either contain one or the other[25,26].

Beyond the strong relationships among them, the hawks seem affiliated to adaptins/coatomers (clusters 4,30) and a family of HEAT repeats with diverse functions, harboring for instance PP2A regulatory subunits (cluster 2, connectivity cluster score > 0.9; Supplementary Table 6), as was suggested before[20]. Adaptins/coatomers comprise a protein family involved in eukaryotes' vesicular transport and the endomembrane system[33]. Whereas we do not find direct links of hawks to prokaryotic proteins, adaptins are found in Asgard archaea[5], which would suggest hawks to have emerged and diverged from the adaptin family during eukaryogenesis.

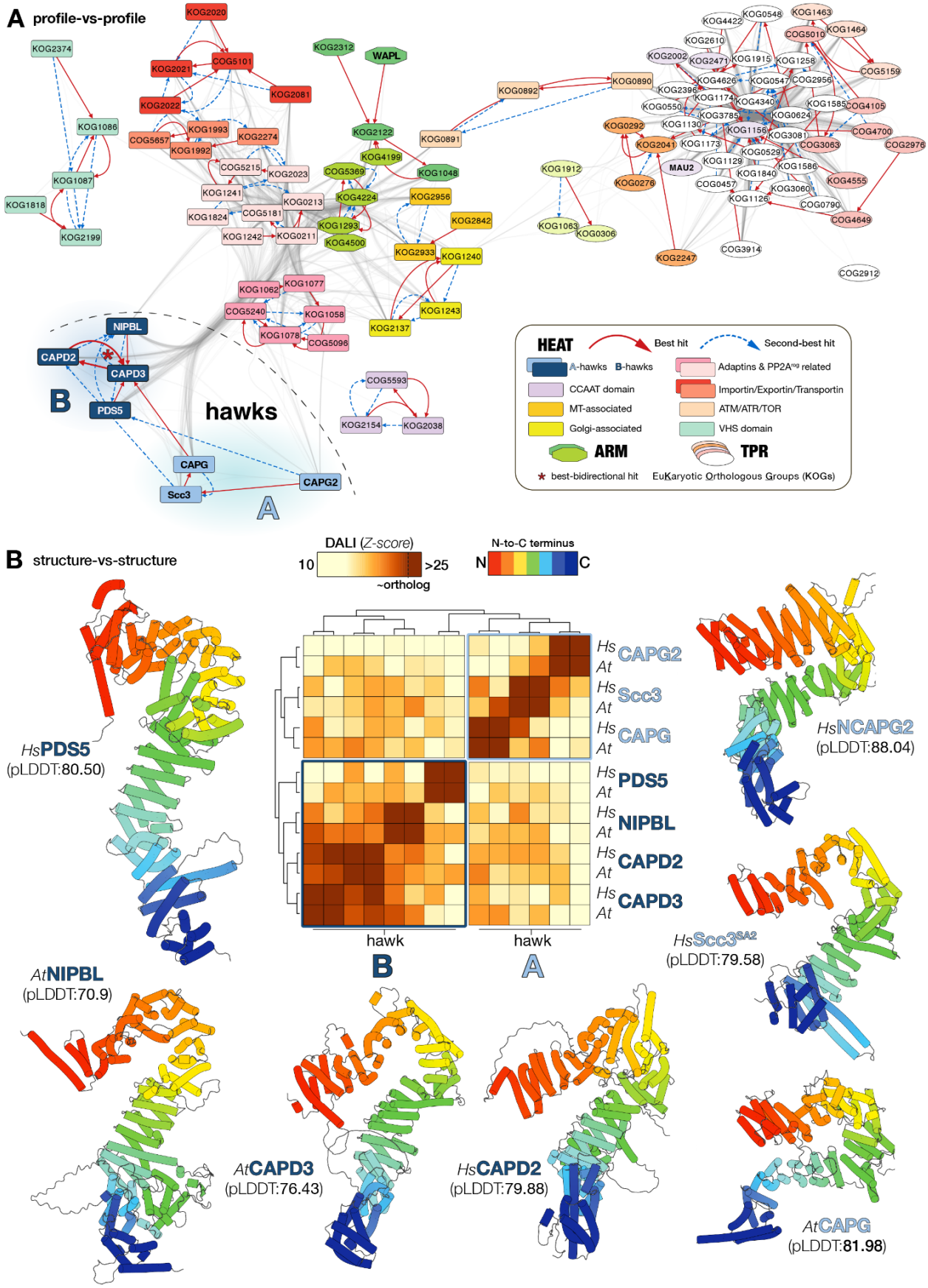

A profile-vs-profile
WAPL
NIPBL
CAPD2
CAPD3
PDS5
CAPG
CAPG2
Scc3
B
A
hawks
HEAT
Best hit
Second-best hit
A-hawks
B-hawks
Adaptins & PP2A^mg related
CCAAT domain
Importin/Exportin/Transportin
MT-associated
ATM/ATR/TOR
Golgi-associated
VHS domain
ARM
TPR
best-bidirectional hit
EuKaryotic Orthologous Groups (KOGs)
B structure-vs-structure
DALI (Z-score)
10
>25
~ortholog
N-to-C terminus
N
C
Hs CAPG2
At
Hs Scc3
At
Hs CAPG
At
Hs PDS5
At
Hs NIPBL
At
Hs CAPD2
At
Hs CAPD3
At
hawk
B
hawk
A
HsPDS5
(pLDDT:80.50)
HsNCAPG2
(pLDDT:88.04)
HsScc3SA2
(pLDDT:79.58)
AtNIPBL
(pLDDT:70.9)
AtCAPD3
(pLDDT:76.43)
HsCAPD2
(pLDDT:79.88)
AtCAPG
(pLDDT:81.98)

Figure 6. Profile HMM network analysis and 3D protein structure comparisons reveal an evolutionary split between A- and B-type hawks. A. Network graph of profile-vs-profile-based network analysis. The initial query profiles were those of hawk proteins, WAPL and MAU2, which were searched against the COG/KOG database. All profile hits identified in the initial search were subsequently taken as query, creating a bidirectional graph. Displayed edges are filtered to show only significant hits (E-value <= 1E-3, alignment length >= 100, which corresponds to ~2 HEAT repeat units). Nodes are colored based on MCL cluster identity (Methods). B. Predicted structures of hawk proteins of *Homo sapiens* (*Hs*) and *Arabidopsis thaliana* (*At*) (Methods). Pairwise alignment Z-scores as computed by Dali [34] are represented in the heatmap, generated with ClustVis [35].

Discussion

An evolutionary scenario for the emergence of eukaryotic SMC complexes

Our phylogenetic analyses of SMC proteins, kleisins and kites point to an ancient duplication of the entire SMC complex components in an archaeal ancestor (Figure 7). We hypothesize that the resulting duplicated SMC proteins formed their own separate homodimer, rather than combining into a single heterodimeric SMC complex. This idea is supported by the genomic co-localization of the candidate subunits of each complex in current-day archaeal genomes (Supplementary Text, Supplementary Figure 3).

From the outset of two paralogous complexes in archaea, we propose that during eukaryogenesis, each complex experienced a duplication and subsequent subfunctionalization of its SMC protein, resulting in the formation of heterodimers constituted by a κ-SMC and a ν-SMC (Figure 7). In the condensin/cohesin ancestor, a kite protein was replaced by a hawk protein. This displacement co-occurred with an increase in the length of the kleisin, as condensin and cohesin have longer kleisin proteins compared to the SMC5/6 complex and to prokaryotic SMC complexes[20,32]. Whereas we were not able to carry out a phylogenetic analysis for hawk proteins due to their extensive divergence[20], our network analysis indicates that the first hawk formed a homodimer, but probably soon duplicated and formed a heterodimer, consisting of an A- and a B-type hawk[32]. This SMC complex then duplicated to give rise to condensin and cohesin. Condensin underwent a partial duplication, affecting only the kleisin and hawks, leading to the formation of condensin I and II. Additionally, the kleisin and one hawk duplicated in cohesin. In the SMC5/6 ancestor, the kite protein duplicated, an event that may have preceded the divergence of eukaryotes from Asgard archaea (Figure 5C). In addition to these duplications, the SMC5/6 complex and cohesin recruited additional subunits and regulators, such as Nse2 and Nse5,6, and WAPL, Securin and Sororin, respectively. Such additions were complemented by existing members that acquired new protein domains, such as Nse1 gaining a RING-finger E3 ubiquitin ligase domain[36], absent in Nse3 and archaeal kites. This domain is likely a eukaryotic innovation, potentially enabling the SMC5/6 complex to ubiquitinate proteins[37]. Combined, these processes endowed LECA with the four SMC complexes found in present-day eukaryotes. Eukaryotic SMC complexes highlight the importance of gene duplications in developing eukaryotic complexity[38,39], involving at least two whole-complex duplications, one partial complex duplication and multiple intra-complex duplications.

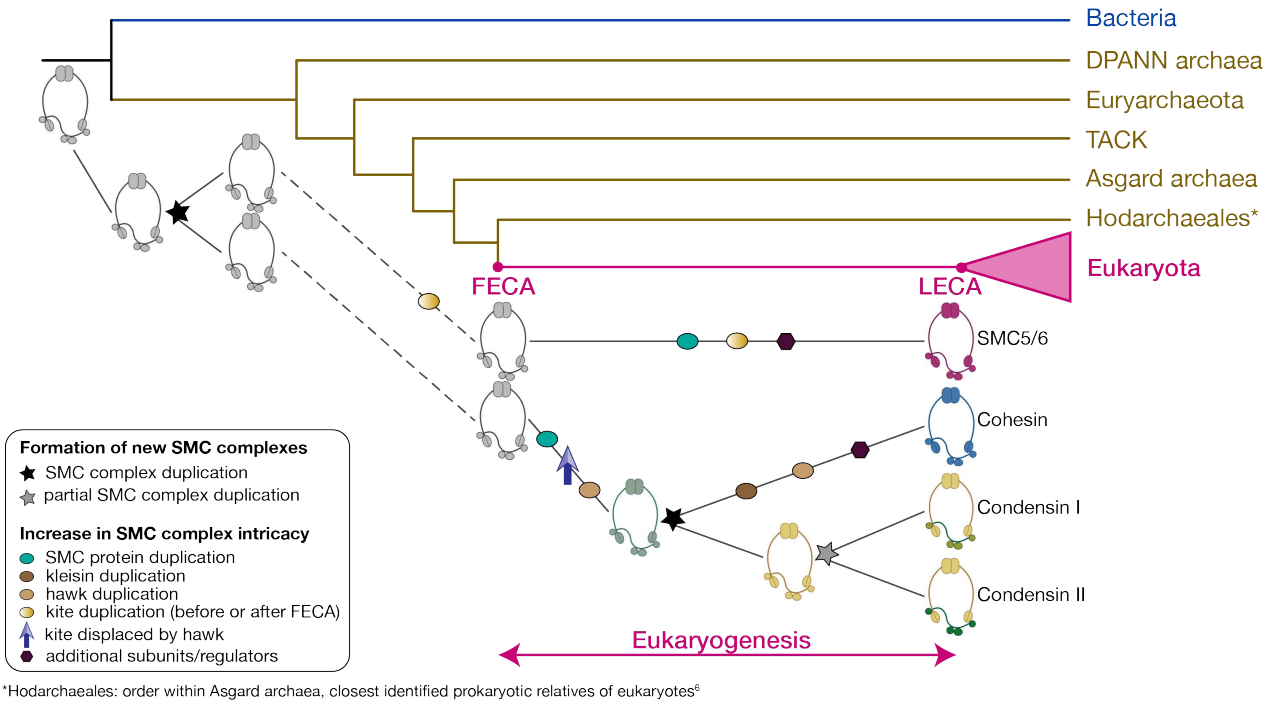


Figure 7. Evolutionary scenario for the origins of the four eukaryotic SMC complexes between the First Eukaryotic Common Ancestor (FECA) and LECA, referred to as 'eukaryogenesis'. FECA possessed two distinct SMC complexes, a legacy of an early archaeal duplication event of the SMC complex (dashed lines). One of these complexes underwent a duplication of both its SMC and kite proteins, leading to the evolution of the SMC5/6 complex. Note that the timing of the kite duplication is unresolved: it might either have occurred before FECA (it would be shared with Asgard archaea, Figure 5C) or after FECA. The SMC duplication transformed the original SMC protein homodimer within the complex into a heterodimer, comprising a κ-SMC and a ν-SMC. The second complex also experienced a duplication of its SMC protein. However, in this instance, the kite subunits were replaced by hawk proteins. This SMC complex underwent an additional duplication, resulting in the emergence of cohesin and condensin. Subsequently, the condensin complex underwent a partial duplication, specifically affecting its non-SMC proteins, namely kleisin and hawks. Finally, within the ancestral condensin/cohesin, the hawk proteins also duplicated. The timing of the hawk duplications is speculative, as we were not able to conduct a phylogenetic analysis for this family. The composition of SMC complexes is illustrated in Figure 1 (inset). Created with BioRender.com.

Outlook: towards evolutionary cell biology of SMC complexes

SMC complexes represent an ancient system for organizing DNA. Like many cellular systems, they have undergone significant expansion and diversification during eukaryogenesis (Figure 7). Moreover, extant eukaryotes differ in their SMC complex repertoires, as recent advances in generating sequencing data for a broader eukaryotic diversity enabled us to reveal. We here provide an evolutionary perspective on eukaryotic SMC complexes, as well as a valuable resource for further exploring their composition, mechanisms, and functions in both model and non-model organisms.

For example, the distribution patterns of condensin I and II across eukaryotes raise questions about their functional differences. What roles does condensin I play in species lacking condensin II? Do they fulfill any non-mitotic functions, such as condensin I's proposed

involvement in TAD formation in *Caenorhabditis elegans*[40,41]? Do species lacking condensin II exhibit a Rabl-like nuclear genome architecture rather than a chromosome territorial structure? This question is particularly relevant to experimentally investigate in fungi, where we identified condensin II in certain non-dikaryan taxa, and in the moss *Physcomitrium patens*, which does not appear to have condensin II. Exploring lineages where SMC5/6 has been recently lost, like in the nematode *S. ratti*, could also enhance our understanding of its dispensability. In addition to raising questions about the implications of such presences and absences, our ortholog sets allow for an in-depth examination of protein sequences, which may reveal new sequence motifs mediating functions, for example through protein-protein interactions.

Our results on the archaeal origins of eukaryotic SMC complexes suggest that various archaea, including species from the TACK and Asgard archaea, possess both a condensin/cohesin-related and an SMC5/6-related complex. This hypothesis could be experimentally tested in various current-day archaea, by assessing if they indeed have two different SMC complexes in vivo. In addition to validating this hypothesis, experimentation in such archaea might assist us in deducing the ancestral functions of condensin and cohesin, and of SMC5/6. In turn, such investigations might help us to understand how eukaryotic genome organization emerged, since this organization is in part shaped by eukaryotes' SMC complexes, yet differs strongly from prokaryotic genome organization[42]. Characterizing the SMC5/6-related archaeal complex is particularly interesting since SMC5/6's (canonical) function in eukaryotes is still enigmatic.

In summary, our work lays the groundwork for a deeper understanding of SMC complexes across the eukaryotic tree of life and their archaeal relatives. As experimental cell biology in non-model species advances, we anticipate gaining richer insights into the evolution and functional diversity of SMC complexes.

Methods

Eukaryotic dataset assembly

To determine the presences and absences of eukaryotic SMC complex subunits across the eukaryotic tree of life, and to reconstruct their evolutionary histories, we used a eukaryotic dataset that we compiled previously[43]. This dataset contains the protein sequences of phylogenetically diverse eukaryotes, and model organisms. For this project, we modified the dataset as described before[44]. An overview of the included species can be found in Supplementary Table 1, where we also report the proteome completeness scores derived from BUSCO[45]. While we attempted to remove redundancy from protein sets, some sets still contain some highly similar sequences, very likely representing isoforms encoded by the same gene. Therefore, some apparent species-specific paralogs in the identified orthologs for a given protein might not correspond to different loci, but to a single locus.

Ortholog detection in eukaryotes

Protein selection and ortholog definition

As the subjects of our study, we selected the subunits of condensin I, condensin II, cohesin and SMC5/6. Additionally, we selected proteins regulating these complexes, provided that they are specific (i.e., no general cell cycle regulators like PP2A, CDK1, and Aurora kinases). The selected proteins should have been characterized as a complex component in at least one organism, predominantly *H. sapiens*, *S. cerevisiae* and *S. pombe*[22,46], and also *T. thermophila*[10,16,47,48], *P. patens*[49,50], *A. thaliana*[28,51] and *C. merolae*[18].

For each selected protein, our objective was to detect its orthologs among the protein sequences of the species in our eukaryotic dataset. Here, ‘orthologs’ are defined as descendants from a single ancestral protein in LECA, or in a more recent lineage, if the protein is taxonomically restricted to a particular eukaryotic clade. As a result, groups of orthologs include paralogs resulting from recent duplications, that is, more recent than LECA. For example, PDS5A,B are both members of the group of PDS5 orthologs.

Ortholog calling: two global approaches

Each protein required its own, tailored approach to ortholog calling, of which the details can be found in Supplementary Table 2 (sheet 'Orthologs').

Our approaches can be globally described by two different strategies. First, for the SMC proteins SMC1-6, we searched for homologs in our eukaryotic dataset with BLASTP[52] using previously identified SMC protein sequences[7]. We combined SMC protein sequences from[7] with our hits and aligned all of them with MAFFT (v7, settings: –auto)[53], trimmed the alignment with BMGE (v1.12, settings: -m BLOSUM30, -b 3 -h 0.7 -g 0.7)[54] and inferred an initial a maximum

likelihood phylogeny using IQ-TREE (v2.0.3, settings: -m LG+F+R5 -alrt 1000)[30]. From this phylogeny, we extracted the sequence belonging to each of SMC1-6, for which we subsequently made protein-specific profile HMMs. We used these HMMs to locally run hmmsearch (hmmer.org) across our eukaryotic dataset. We gathered the hit sequences using a domain bitscore cut-off that was well below the lowest-scoring members of putative orthologs in the initial phylogeny, to ensure all orthologs were included. For these hit sequences, we inferred a final phylogeny from which the orthologs were extracted.

Second, for non-SMC proteins, we derived profile HMMs from PANTHER17.0[55] and EggNOG v6.0[56]. We used these HMMs to search across our eukaryotic dataset, and verified that characterized orthologs and/or functional analogs were hit by it. Then, we performed at least one iteration of the following steps: construct a profile HMM of the hits and search across our eukaryotic dataset, verify the new hits manually, and gather the trusted hits. In the next iteration, the trusted hits served as input for a new profile HMM. As long as new hits were found, we continued this cycle. For highly divergent proteins, we adopted additional strategies. These involved sequence-based strategies, like BLASTP/PSI-BLAST[52,57], online jackhmmer[58] and online HHpred[59], frequently using multiple query sequences. Since sequences typically are less conserved than structures, we also adopted structure-based strategies, which entail searching for similar structures using experimentally determined structures or predicted structures from the AlphaFold Protein Structure Database[60], applying search algorithms Dali[34] and Foldseek[61], also often using multiple query structures. The collected homologs were aligned to infer a maximum likelihood phylogeny (details in Supplementary Table 2, sheet 'Orthologs'), from which the orthologs were extracted. If, because of strong sequence divergence, we expected the orthologous cluster in the phylogeny to be incomplete, we used the orthologs in the cluster to build a profile HMM to again search through our dataset, infer another phylogeny and extract the orthologs, potentially expanded compared to the previous set. Particularly Securin and Sororin, large comprising intrinsically disordered regions, form a challenge for ortholog detection, since they are highly divergent proteins for which no structure-based strategy would work. A detailed description of the approach followed for Securin and Sororin can be found in the Supplementary Text.

For each protein, we provide the following data files in a GitHub repository (https://github.com/jolienvanhooff/SMCevolution):

- a text file containing the ortholog identifiers;
- a FASTA file containing unaligned orthologs;
- the profile HMM used to search for orthologs;
- the FASTA file containing the (curated) multiple sequence alignment for the profile HMM;
- the FASTA file containing the multiple sequence alignment used for phylogenetic inference;
- the logfile of the phylogenetic inference;
- the annotated gene phylogeny.

Phylogenetic and network analysis in eukaryotes and prokaryotes

Eukaryotic and prokaryotic dataset assembly

To study the pre-LECA origins of eukaryotic SMC complex subunits, we assembled two protein datasets: one for eukaryotes and one for prokaryotes. The eukaryotic dataset consists of a subset of the species we used to find orthologs in, which we subsampled to accelerate the analyses and increase the interpretability of the results, while still spanning eukaryotic diversity. The selected 49 species can be found in Supplementary Table 1 (column 'PreLECASubset'). We annotated the orthologs of our proteins of interest, to facilitate the analyses and interpretation. The prokaryotic dataset was derived from GTDB (r207)[62]. We used a subset of the assemblies in GTDB, ultimately including one assembly for each family-level clade. For each family, we selected assemblies that A), are the GTDB representative for a given species, B), have a CheckM quality score (completeness score - 5*contamination score) above 50, and C), have the best CheckM quality score of their family. This resulted in a prokaryotic dataset containing 456 archaea and 3645 bacteria.

Protein selection

We selected proteins that we deemed informative about the deep evolutionary origins of SMC complexes. We thereto applied the following criteria:

1. The protein was present in LECA;
2. The protein was present in an SMC complex in LECA as a permanent constituent, i.e. not as a regulatory or transient subunit;
3. The protein belongs to a family of typical SMC complex subunits, and is not likely to be derived from other cellular machineries.

Based on these criteria we included the following proteins and their families: SMC1-6 (SMC protein family), CAPH, CAPH2, Scc1, Rec8 and Nse4 (kleisin family), CAPG, CAPD2, CAPG2, CAPD3, Scc3, PDS5 and NIPBL (hawk family), Nse1 and Nse3 (kite family), and Nse5 (family unknown) and Nse6 (family unknown). As part of our effort to find the evolutionary origins of the hawk family and Nse5,6, we searched for those of MAU2 and WAPL as well, in our HMM network analysis (see below).

Phylogenetic analysis

We restricted the phylogenetic analysis to the SMC proteins, the kleisins and the kites, since we deemed the hawks, Nse5 and Nse6 too divergent for phylogenetic analysis. An overview of the applied phylogenetic methods for each family can be found in Supplementary Table 2 (sheet 'Homologs'). In general, we used profile HMM of separate orthologous sets (SMC proteins) or, for less conserved proteins, a profile HMM based on their merged alignments (kleisin, kite) to first search across our eukaryotic dataset. With this first search, we confirmed that there are no obvious other eukaryotic homologs beyond the proteins making up the SMC complexes[7]. Then, we searched iteratively across our prokaryotic sequence dataset, and collected significant prokaryotic hits. We retrieved the eukaryotic protein sequences from the ortholog sets, i.e., those belonging to species of the subsampled dataset, and combined these with the prokaryotic

hits. To allow for phylogenetic inference using a relatively complex substitution model (see below), which is computationally more intense, we further downsampled the SMC and kleisin sequences based on sequence identity with CD-HIT (v4.8.1, different identity cutoffs, Supplementary Table 2, sheet 'Homologs')[63]. For kite, such downsampling was not necessary, because for this family we limited the phylogenetic analyses to eukaryotic sequences and their readily discernible homologs in archaea, due to which the total number of sequences was lower. We chose to exclude bacterial kites, as well as archaeal kites that we were not able to recruit with HMM-based sequence searches, since these were likely to be too distinct from the eukaryotic sequences to meaningfully align them, or to model the evolution of in the phylogenetic inference.

The remaining protein sequences were aligned using MAFFT-L-INS-i or MAFFT-DASH[64] (v7.515), or using the PROMALS3D web server[65]. We selected the alignment approach yielding the most positions after trimming, as well as the best phylogeny (discussed below). For PROMALS3D, we obtained predicted protein structures from diverse sequences from the AlphaFold Protein Structure Database[60], and included them as input. We trimmed the resulting multiple sequence alignments (MSAs) with trimAl in the 'gappyout' mode (v1.4.1)[66] and manually cut out regions not well-conserved. We cleaned the MSA from sequences that consisted of over >70% of gaps. Phylogenetic analysis was implemented through inference of a maximum likelihood (ML) phylogeny with IQ-TREE (v.2.0.3)[30], using the site-heterogeneous LG+C60+G substitution model and assessing branch support using ultrafast bootstrap approximation (UFBOOT[31], 1000 replicates). To test the topology of the resulting phylogenies, we inferred another phylogeny with the same model, constraining the eukaryotic sequences to be monophyletic. We compared the original tree (the ML tree or, if it had a higher likelihood, the consensus tree) to the constrained tree to assess if the constrained tree can be rejected using the approximately unbiased (AU) test provided by IQ-TREE.

For all families, we attempted to use PROMALS3D for generating an MSA, because it, in theory, provides a means to improve the alignment of divergent sequences, since the sequence alignment is guided by the alignment of (more conserved) structures. While this worked well for the SMC proteins, it did not for the kleisin and kite families, as subsequent phylogenetic inference failed to recover the individual eukaryotic orthologous groups. Sequence members of these orthologous groups were not monophyletic, while they should have been. Moreover, ultrafast bootstrap support values were low for various key bipartitions. Possibly, these poor signals are resulting from a flawed structural guide alignment, although we did not further examine the potential underlying causes. For kites, we were able to use some structural information by employing MAFFT-DASH, which recruits resolved protein structures from the PDB, based on sequence similarity. Since it does not use predicted protein structures, the number of structures used as guides is low. For kleisins, generating a good MSA turned out particularly challenging. Therefore, we applied MAFFT-L-INS-i to the individual orthologous groups (except for Scc1 and Rec8, who were combined into a single alignment, because they were interspersed in previous phylogenies), and subsequently merged them using the ‘merge’ option, also adding the prokaryotic homologs.

PROMALS3D alignment-based phylogenies were pruned to remove leaves that corresponded to input structures. Furthermore, the SMC phylogeny, in the final visualization in Figure 5, was pruned from four leaves with very long branches (>5 average substitutions per site), and with unclear phylogenetic affiliations, since these might comprise spurious sequences. The clades were annotated according to A), their membership of a eukaryotic orthologous group (eukaryotic sequences) and B, their sisterhood relationship to the eukaryotic orthologous groups (prokaryotic sequences). If their relationship to eukaryotic sequences was not clear, we labeled them as 'unknown' (not necessarily monophyletic), or, in the case of the SpcB-like kites, as 'ScpB-like'. The distributions of each group across archaea are provided in Supplementary Table 3. Note that, for SMC and kleisin proteins, these distributions form an underrepresentation due to the downsampling of the sequences for the phylogenetic analysis, as described above.

To specifically explore the relationships among the condensin/cohesin kleisins (CAPH, CAPH2, Scc1 and Rec8), we inferred an additional, more restricted phylogeny. We collected prokaryotic homologs based on a profile HMM constructed with a MAFFT-L-INS-i alignment of these four proteins. Then, we aligned the eukaryotic and prokaryotic sequences using MAFFT-LINSI and 'merge' as described above, trimmed the alignment with trimAl 'gappyout', removed sequences with >70% gaps, and applied IQ-TREE (substitution model: LG+C60+G, 1000 UFBOOT replicates).

For the phylogenies inferred for the SMC, kleisin and kite proteins, we shared on our GitHub repository (https://github.com/jolienvanhooff/SMCevolution):

- The unprocessed multiple sequence alignment
- The multiple sequence alignment used as input for phylogenetic inference
- The maximum likelihood and the consensus phylogenies
- The logfile from the maximum likelihood phylogenetic inference
- The maximum likelihood phylogeny inferred with constraint topology, i.e. with eukaryotic sequences forced to form a monophyly.
- The reporting file of the topology test.

In addition, we share the unfiltered, uncollapsed, annotated phylogenies corresponding to Figure 5A-C on the iTOL web server[67]. Hyperlinks can be found in the legend of Figure 5.

Constructing the profile-versus-profile HMM network

Reconstructing the evolutionary history of the hawk proteins, Nse5,6, WAPL, and MAU2 proved challenging using classical homology detection and phylogenetic inference methods due to their high sequence divergence, as noted before[20]. To address these difficulties, we employed the profile-versus-profile HMM-based network approach introduced by[20]. Briefly, we constructed a target profile database consisting of the MSAs generated here and combined this with the precomputed COG_KOG_v1.0 database (downloaded from http://ftp.tuebingen.mpg.de/pub/protevo/toolkit/databases/hhsuite_dbs/COG_KOG.tar.gz, last accessed Jan 17, 2023). We queried our manually-curated MSAs against this target database using the HHsearch algorithm from the HH-suite package[68] (v3.3.0) with default parameters as implemented in the MPI Bioinformatics Toolkit[68,69]. Following this initial search, all KOG/COG

profiles that were hit by a query profile (E<1000) in this first step were then used to complete reciprocal searches on the same database. The resultant graph (Supplementary Table 5) was visualized using Cytoscape[70]. Here, we first imposed thresholds for the E-value (<1e-3) and alignment length (>=100), to keep only the edges that imply meaningful significant similarity between profiles. Then, Markov Chain Algorithm (MCL) clustering was performed with an inflation parameter *I* = 3 using 1/E-value as array source as implemented in the clusterMaker2 package[71]. Finally, we used Cytoscape's Prefuse Force Directed Layout with default settings, except for iterations = 1000, default node mass = 10, and we set the 'Force deterministic layouts' to 'on' and did not partition the graph before applying the layout. We also calculated the 'connectivity' between clusters. Connectivity is defined as the number of observed edges going from cluster 1 to cluster 2 (E-value=<1e-3 and alignment length>=100) over the total number of possible edges (Supplementary Table 6).

Hawk structural alignment

Predicted 3D structures of hawks from both *A. thaliana and H. sapiens* were directly downloaded from the AlphaFold repository (alphafold.ebi.ac.uk/, last accessed 23 December 2023). Structures were manually checked in Pymol and N- or C-terminal disordered regions and/or other domains that do not pertain to the core HEAT-repeat region were removed (see for coordinates and uniprot identifiers Supplementary Table 7). The DALI webserver[34] (Holm et al., 2006) was used for structural comparisons amongst the hawks, using the all-against-all function (ekhidna2.biocenter.helsinki.fi/dali). Z-scores were average clustered with the pearson coefficient as a distance measure (Supplementary Table 7). A Pymol session is included in the Supplementary Data.

Nse5,6 alpha-solenoid structure prediction and alignment

To predict the structures of Nse5,6 orthologs, we used the AlphaFold2 software[72] through the local ColabFold implementation[73]. We used alignments produced by MAFFT-L-INS-i of our curated sets of orthologs as input to ensure deep alignment information to produce high-quality structure predictions. Subsequently, the alpha-solenoids were isolated from the full-length structures using the pdb_tools package[74]. The start and end coordinates of the alpha-solenoids of each full-length structure were determined by aligning them to manually-curated human Nse5 or Nse6 alpha-solenoids using Foldseek[61]. All isolated Nse5 and Nse6 alpha-solenoids were then comprehensively aligned in a pairwise fashion using TMalign[75], where a TM-score of >=0.5 was considered the threshold for significance[76].

Visualization

The simplified species tree (Figures 1,7), the distributions of SMC complex subunits across eukaryotes (Figures 2-4) and the protein family phylogenies (Figure 5) were drawn using iTOL[67]. Phylogenies were edited and iTOL input datasets were generated using ETE (v3)[77] and Biopython[78]. Structures for Figure 6 were visualized and edited using PyMOL[79]. Structures for

Supplementary Figure 4 were visualized using UCSF Chimera[80]. The heatmap in Figure 6B was visualized using R[81]. All final figures were prepared using Adobe Illustrator.

Acknowledgements

We kindly thank the following colleagues:

- Miguel Andrade-Navarro (Johannes-Gutenberg University), for his helpful advice and expert assessment on the classification of the alpha solenoid folds of Nse5 and Nse6;
- Jason Stajich (University of California Riverside) for sharing the predicted protein sequences of *Gonapodya* sp. JEL0774;
- Bastiaan de Potter, Laura van Rooijen, Carlos Santana-Molina and Berend Snel (Utrecht University) for collaborating in compiling the eukaryotic proteome dataset;
- Benjamin Rowland (Netherlands Cancer Institute) for his constructive feedback on a preliminary version of this manuscript.

This work was supported by the European Research Council (ERC Starting Grant 2018, grant number 803151, to L.E.) and the Dutch Research Council (NWO, VI.Veni.212.099 to J.J.E.H. and VI.Veni.202.223 to E.C.T.). The authors declare no competing financial interests.

Author contributions

Conceptualization, J.J.E.H., E.C.T., and L.E.; data curation, J.J.E.H., M.W.D.R., and E.C.T.; analysis, J.J.E.H, M.W.D.R., and E.C.T.; methodology, J.J.E.H, M.W.D.R, E.C.T., and L.E.; project administration, J.J.E.H.; resources, J.J.E.H, E.C.T, L.E.; supervision, J.J.E.H, E.C.T., and L.E; validation, J.J.E.H., M.W.D.R., and E.C.T.; visualization, J.J.E.H., M.W.D.R., and E.C.T.; writing – original draft, J.J.E.H. and L.E.; writing – reviewing and editing; J.J.E.H, M.W.D.R, E.C.T., and L.E.

Supplementary Material

Supplementary Information, including Supplementary Text, Supplementary Figures 1-4, Supplementary Table 4, and information on Supplementary Tables 1-3,5-7 and on Supplementary Data.

References

1. Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. Sci Adv *5*, eaaw1668. 10.1126/sciadv.aaw1668.

2. Davidson, I.F., and Peters, J.-M. (2021). Genome folding through loop extrusion by SMC complexes. Nat. Rev. Mol. Cell Biol. *22*, 445–464. 10.1038/s41580-021-00349-7.

3. Hassler, M., Shaltiel, I.A., and Haering, C.H. (2018). Towards a Unified Model of SMC Complex Function. Curr. Biol. *28*, R1266–R1281. 10.1016/j.cub.2018.08.034.

4. Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., van Eijk, R., Schleper, C., Guy, L., and Ettema, T.J.G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature *521*, 173. 10.1038/nature14447.

5. Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature *541*, 353–358. 10.1038/nature21031.

6. Eme, L., Tamarit, D., Caceres, E.F., Stairs, C.W., De Anda, V., Schön, M.E., Seitz, K.W., Dombrowski, N., Lewis, W.H., Homa, F., et al. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Nature *618*, 992–999. 10.1038/s41586-023-06186-2.

7. Yoshinaga, M., and Inagaki, Y. (2021). Ubiquity and Origins of Structural Maintenance of Chromosomes (SMC) Proteins in Eukaryotes. Genome Biol. Evol. *13*. 10.1093/gbe/evab256.

8. King, T.D., Leonard, C.J., Cooper, J.C., Nguyen, S., Joyce, E.F., and Phadnis, N. (2019). Recurrent Losses and Rapid Evolution of the Condensin II Complex in Insects. Mol. Biol. Evol. *36*, 2195–2204. 10.1093/molbev/msz140.

9. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeño Cacciatore, Á., Contessoto, V.G., van Heesbeen, R.G.H.P., van den Broek, B., et al. (2021). 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. Science *372*, 984–989. 10.1126/science.abe2218.

10. Howard-Till, R., Tian, M., and Loidl, J. (2019). A specialized condensin complex participates in somatic nuclear maturation in Tetrahymena thermophila. Mol. Biol. Cell *30*, 1326–1338. 10.1091/mbc.E18-08-0487.

11. Hirano, T. (2016). Condensin-Based Chromosome Organization from Bacteria to Vertebrates. Cell *164*, 847–857. 10.1016/j.cell.2016.01.033.

12. Bauer, C.R., Hartl, T.A., and Bosco, G. (2012). Condensin II promotes the formation of chromosome territories by inducing axial compaction of polyploid interphase chromosomes. PLoS Genet. *8*, e1002873. 10.1371/journal.pgen.1002873.

13. Ay, F., Bunnik, E.M., Varoquaux, N., Bol, S.M., Prudhomme, J., Vert, J.-P., Noble, W.S., and Le Roch, K.G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. *24*, 974–988. 10.1101/gr.169417.113.

14. Bunnik, E.M., Venkat, A., Shao, J., McGovern, K.E., Batugedara, G., Worth, D., Prudhomme, J., Lapp, S.A., Andolina, C., Ross, L.S., et al. (2019). Comparative 3D genome organization in apicomplexan parasites. Proc. Natl. Acad. Sci. U. S. A. *116*, 3183–3192. 10.1073/pnas.1810815116.

15. Hirano, T. (2012). Condensins: universal organizers of chromosomes with diverse functions.

Genes Dev. *26*, 1659–1678. 10.1101/gad.194746.112.

16. Howard-Till, R., and Loidl, J. (2018). Condensins promote chromosome individualization and segregation during mitosis, meiosis, and amitosis in Tetrahymena thermophila. Mol. Biol. Cell *29*, 466–478. 10.1091/mbc.E17-07-0451.

17. Kalitsis, P., Zhang, T., Marshall, K.M., Nielsen, C.F., and Hudson, D.F. (2017). Condensin, master organizer of the genome. Chromosome Res. *25*, 61–76. 10.1007/s10577-017-9553-0.

18. Fujiwara, T., Tanaka, K., Kuroiwa, T., and Hirano, T. (2013). Spatiotemporal dynamics of condensins I and II: evolutionary insights from the primitive red alga Cyanidioschyzon merolae. Mol. Biol. Cell *24*, 2515–2527. 10.1091/mbc.E13-04-0208.

19. Peng, X.P., and Zhao, X. (2023). The multi-functional Smc5/6 complex in genome protection and disease. Nat. Struct. Mol. Biol. *30*, 724–734. 10.1038/s41594-023-01015-6.

20. Wells, J.N., Gligoris, T.G., Nasmyth, K.A., and Marsh, J.A. (2017). Evolution of condensin and cohesin complexes driven by replacement of Kite by Hawk proteins. Curr. Biol. *27*, R17–R18. 10.1016/j.cub.2016.11.050.

21. Oravcová, M., Nie, M., Zilio, N., Maeda, S., Jami-Alahmadi, Y., Lazzerini-Denchi, E., Wohlschlegel, J.A., Ulrich, H.D., Otomo, T., and Boddy, M.N. (2022). The Nse5/6-like SIMC1-SLF2 complex localizes SMC5/6 to viral replication centers. Elife *11*. 10.7554/eLife.79676.

22. Diaz, M., and Pecinka, A. (2018). Scaffolding for Repair: Understanding Molecular Functions of the SMC5/6 Complex. Genes *9*. 10.3390/genes9010036.

23. Adamus, M., Lelkes, E., Potesil, D., Ganji, S.R., Kolesar, P., Zabrady, K., Zdrahal, Z., and Palecek, J.J. (2020). Molecular Insights into the Architecture of the Human SMC5/6 Complex. J. Mol. Biol. *432*, 3820–3837. 10.1016/j.jmb.2020.04.024.

24. Patil, S., Moeys, S., von Dassow, P., Huysman, M.J.J., Mapleson, D., De Veylder, L., Sanges, R., Vyverman, W., Montresor, M., and Ferrante, M.I. (2015). Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta. BMC Genomics *16*, 930. 10.1186/s12864-015-1983-5.

25. Alonso-Gil, D., and Losada, A. (2023). NIPBL and cohesin: new take on a classic tale. Trends Cell Biol. 10.1016/j.tcb.2023.03.006.

26. Petela, N.J., Gligoris, T.G., Metson, J., Lee, B.-G., Voulgaris, M., Hu, B., Kikuchi, S., Chapard, C., Chen, W., Rajendra, E., et al. (2018). Scc2 Is a Potent Activator of Cohesin's ATPase that Promotes Loading by Binding Scc1 without Pds5. Mol. Cell *70*, 1134–1148.e7. 10.1016/j.molcel.2018.05.022.

27. Eme, L., Trilles, A., Moreira, D., and Brochier-Armanet, C. (2011). The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes. BMC Evol. Biol. *11*, 265. 10.1186/1471-2148-11-265.

28. Mota, I.P., Galova, M., Schleiffer, A., Nguyentan, T.-T., Kovacikova, I., Nishiyama, T., Gregan, J., Peters, J.-M., and Schlogelhofer, P. (2022). SORORIN is an evolutionary conserved antagonist of WAPL. bioRxiv, 2022.10.24.513534. 10.1101/2022.10.24.513534.

29. Prokopchuk, G., Butenko, A., Dacks, J.B., Speijer, D., Field, M.C., and Lukeš, J. (2023). Lessons from the deep: mechanisms behind diversification of eukaryotic protein complexes. Biol. Rev. Camb. Philos. Soc. 10.1111/brv.12988.

30. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534. 10.1093/molbev/msaa015.

31. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522. 10.1093/molbev/msx281.

32. Oldenkamp, R., and Rowland, B.D. (2022). A walk through the SMC cycle: From catching DNAs to shaping the genome. Mol. Cell *82*, 1616–1630. 10.1016/j.molcel.2022.04.006.

33. Rout, M.P., and Field, M.C. (2017). The Evolution of Organellar Coat Complexes and Organization of the Eukaryotic Cell. Annu. Rev. Biochem. *86*, 637–657. 10.1146/annurev-biochem-061516-044643.

34. Holm, L. (2022). Dali server: structural unification of protein families. Nucleic Acids Res. *50*, W210–W215. 10.1093/nar/gkac387.

35. Metsalu, T., and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. *43*, W566–W570. 10.1093/nar/gkv468.

36. Jo, A., Li, S., Shin, J.W., Zhao, X., and Cho, Y. (2021). Structure Basis for Shaping the Nse4 Protein by the Nse1 and Nse3 Dimer within the Smc5/6 Complex. J. Mol. Biol. *433*, 166910. 10.1016/j.jmb.2021.166910.

37. Ibars, E., Codina-Fabra, J., Bellí, G., Casas, C., Tarrés, M., Solé-Soler, R., Lorite, N.P., Ximénez-Embún, P., Muñoz, J., Colomina, N., et al. (2023). Ubiquitin proteomics identifies RNA polymerase I as a target of the Smc5/6 complex. Cell Rep. *42*, 112463. 10.1016/j.celrep.2023.112463.

38. Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. (2005). Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res. *33*, 4626–4638. 10.1093/nar/gki775.

39. Vosseberg, J., van Hooff, J.J.E., Marcet-Houben, M., van Vlimmeren, A., van Wijk, L.M., Gabaldón, T., and Snel, B. (2021). Timing the origin of eukaryotic cellular complexity with ancient duplications. Nat Ecol Evol *5*, 92–100. 10.1038/s41559-020-01320-z.

40. Das, M., Semple, J.I., Volodkina, V., Haemmerli, A., Scotton, J., Gitchev, T., Annan, A., Statzer, C., Ewald, C.Y., Mozziconacci, J., et al. (2022). Condensin I folds the C. elegans genome. bioRxiv, 2022.06.14.495661. 10.1101/2022.06.14.495661.

41. Kim, J., Jimenez, D.S., Ragipani, B., Zhang, B., Street, L.A., Kramer, M., Albritton, S.E.,

Winterkorn, L.H., Morao, A.K., and Ercan, S. (2022). Condensin DC loads and spreads from recruitment sites to create loop-anchored TADs in C. elegans. Elife *11*. 10.7554/eLife.68745.

42. van Hooff, J.J.E. (2023). Towards unraveling the origins of eukaryotic nuclear genome organization. Trends Cell Biol. *33*, 820–823. 10.1016/j.tcb.2023.07.008.

43. de Potter, B., Raas, M.W.D., Seidl, M.F., Peter Verrijzer, C., and Snel, B. (2023). Uncoupled evolution of the Polycomb system and deep origin of non-canonical PRC1. bioRxiv, 2023.04.04.535607. 10.1101/2023.04.04.535607.

44. Stok, C., Tsaridou, S., van den Tempel, N., Everts, M., Wierenga, E., Bakker, F.J., Kok, Y., Alves, I.T., Jae, L.T., Raas, M.W.D., et al. (2023). FIRRM/C1orf112 is synthetic lethal with PICH and mediates RAD51 dynamics. Cell Rep. *42*, 112668. 10.1016/j.celrep.2023.112668.

45. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. *38*, 4647–4654. 10.1093/molbev/msab199.

46. Haering, C.H., and Gruber, S. (2016). SnapShot: SMC Protein Complexes Part I. Cell *164*, 326–326.e1. 10.1016/j.cell.2015.12.026.

47. Howard-Till, R.A., Lukaszewicz, A., Novatchkova, M., and Loidl, J. (2013). A single cohesin complex performs mitotic and meiotic functions in the protist tetrahymena. PLoS Genet. *9*, e1003418. 10.1371/journal.pgen.1003418.

48. Ali, E.I., Loidl, J., and Howard-Till, R.A. (2018). A streamlined cohesin apparatus is sufficient for mitosis and meiosis in the protist Tetrahymena. Chromosoma *127*, 421–435. 10.1007/s00412-018-0673-x.

49. Holá, M., Vágnerová, R., and Angelis, K.J. (2021). Kleisin NSE4 of the SMC5/6 complex is necessary for DNA double strand break repair, but not for recovery from DNA damage in Physcomitrella (Physcomitrium patens). Plant Mol. Biol. *107*, 355–364. 10.1007/s11103-020-01115-7.

50. Lelkes, E., Jemelková, J., Holá, M., Štefanovie, B., Kolesár, P., Vágnerová, R., Tomaštíková, E.D., Pecinka, A., Angelis, K.J., and Paleček, J.J. (2023). Characterization of the conserved features of the NSE6 subunit of the Physcomitrium patens SMC5/6 complex. Plant J. 10.1111/tpj.16282.

51. Yang, C., Hamamura, Y., Sofroni, K., Böwer, F., Stolze, S.C., Nakagami, H., and Schnittger, A. (2019). SWITCH 1/DYAD is a WINGS APART-LIKE antagonist that maintains sister chromatid cohesion in meiosis. Nat. Commun. *10*, 1–15. 10.1038/s41467-019-09759-w.

52. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. 10.1016/s0022-2836(05)80360-2.

53. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. 10.1093/molbev/mst010.

54. Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. *10*, 210. 10.1186/1471-2148-10-210.

55. Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., and Mi, H. (2022). PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. *31*, 8–22. 10.1002/pro.4218.

56. Hernández-Plaza, A., Szklarczyk, D., Botas, J., Cantalapiedra, C.P., Giner-Lamia, J., Mende, D.R., Kirsch, R., Rattei, T., Letunic, I., Jensen, L.J., et al. (2022). eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res. 10.1093/nar/gkac1022.

57. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. 10.1093/nar/25.17.3389.

58. Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. (2015). HMMER web server: 2015 update. Nucleic Acids Res. *43*, W30–W38. 10.1093/nar/gkv397.

59. Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., and Alva, V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr. Protoc. Bioinformatics *72*, e108. 10.1002/cpbi.108.

60. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. 10.1093/nar/gkab1061.

61. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. 10.1038/s41587-023-01773-0.

62. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. 10.1093/nar/gkab776.

63. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. 10.1093/bioinformatics/bts565.

64. Rozewicki, J., Li, S., Amada, K.M., Standley, D.M., and Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. *47*, W5–W10. 10.1093/nar/gkz342.

65. Pei, J., and Grishin, N.V. (2014). PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. Methods Mol. Biol. *1079*, 263–271. 10.1007/978-1-62703-646-7_17.

66. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for

automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. 10.1093/bioinformatics/btp348.

67. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. 10.1093/nar/gkab301.

68. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 473. 10.1186/s12859-019-3019-7.

69. Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol. *430*, 2237–2243. 10.1016/j.jmb.2017.12.007.

70. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

71. Morris, J.H., Apeltsin, L., Newman, A.M., Baumbach, J., Wittkop, T., Su, G., Bader, G.D., and Ferrin, T.E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics *12*, 436. 10.1186/1471-2105-12-436.

72. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. 10.1038/s41586-021-03819-2.

73. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. 10.1038/s41592-022-01488-1.

74. Rodrigues, J.P.G.L.M., Teixeira, J.M.C., Trellet, M., and Bonvin, A.M.J.J. (2018). pdb-tools: a swiss army knife for molecular structures. F1000Res. *7*, 1961. 10.12688/f1000research.17456.1.

75. Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309. 10.1093/nar/gki524.

76. Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895. 10.1093/bioinformatics/btq066.

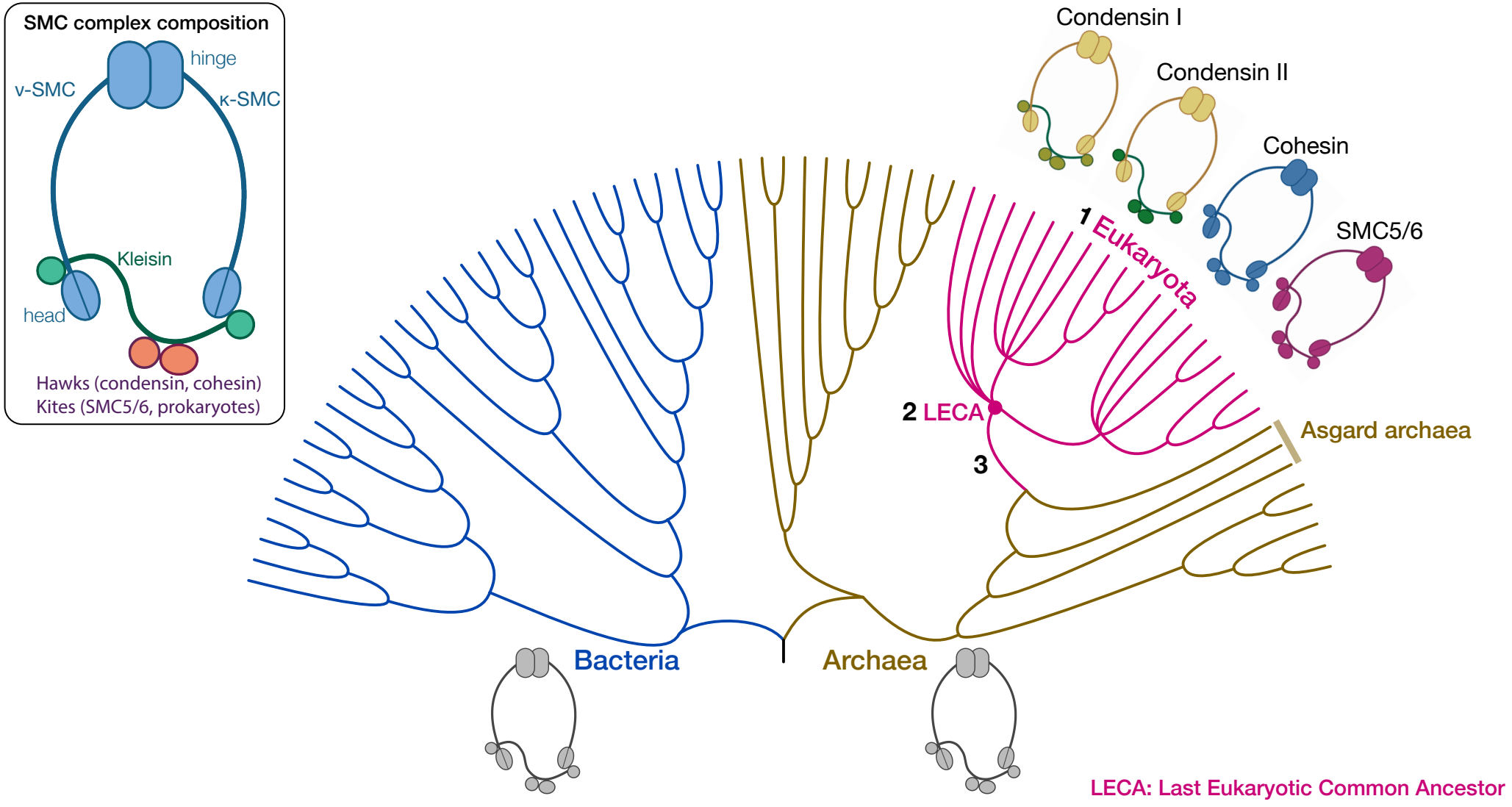
77. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. *33*, 1635–1638. 10.1093/molbev/msw046.

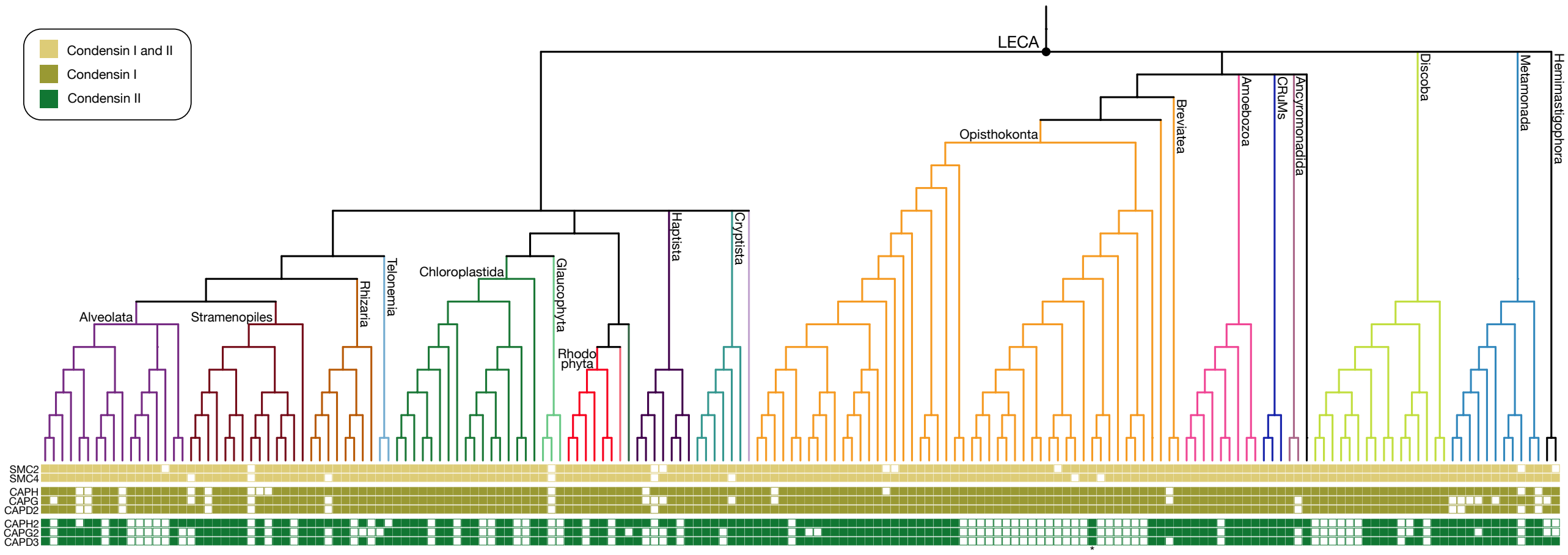
78. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. 10.1093/bioinformatics/btp163.

79. Schrödinger, LLC The PyMOL Molecular Graphics System, Version 2.0.

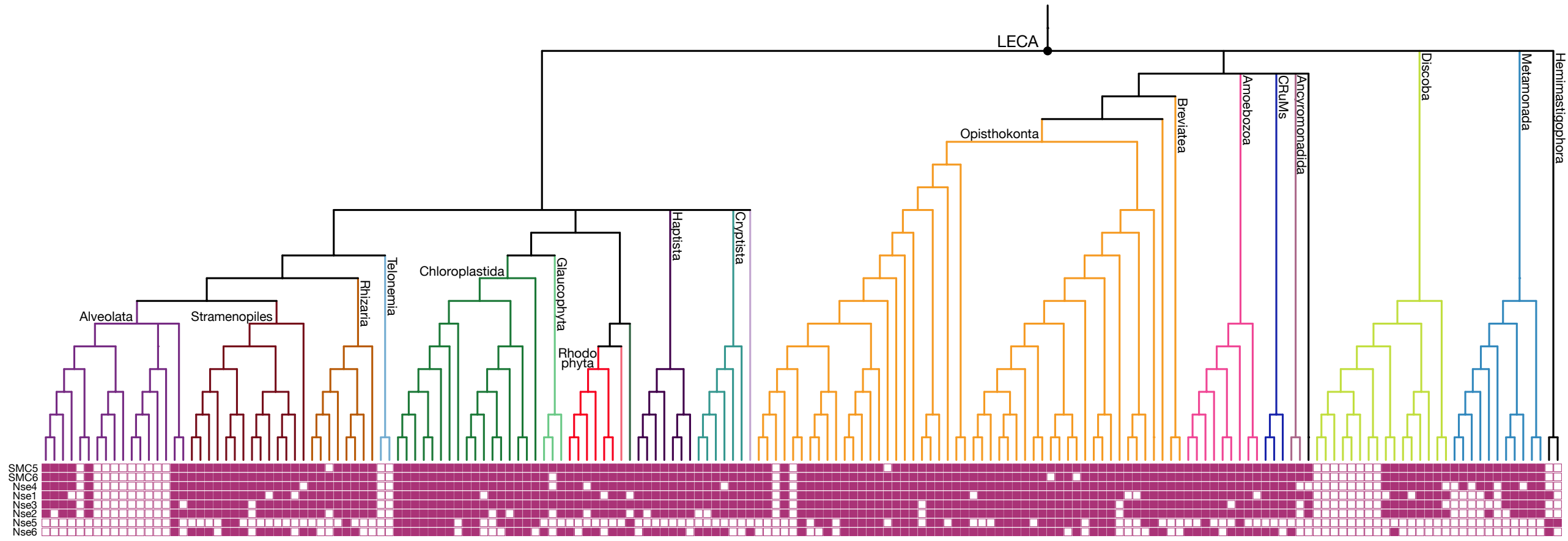
80. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. 10.1002/jcc.20084.

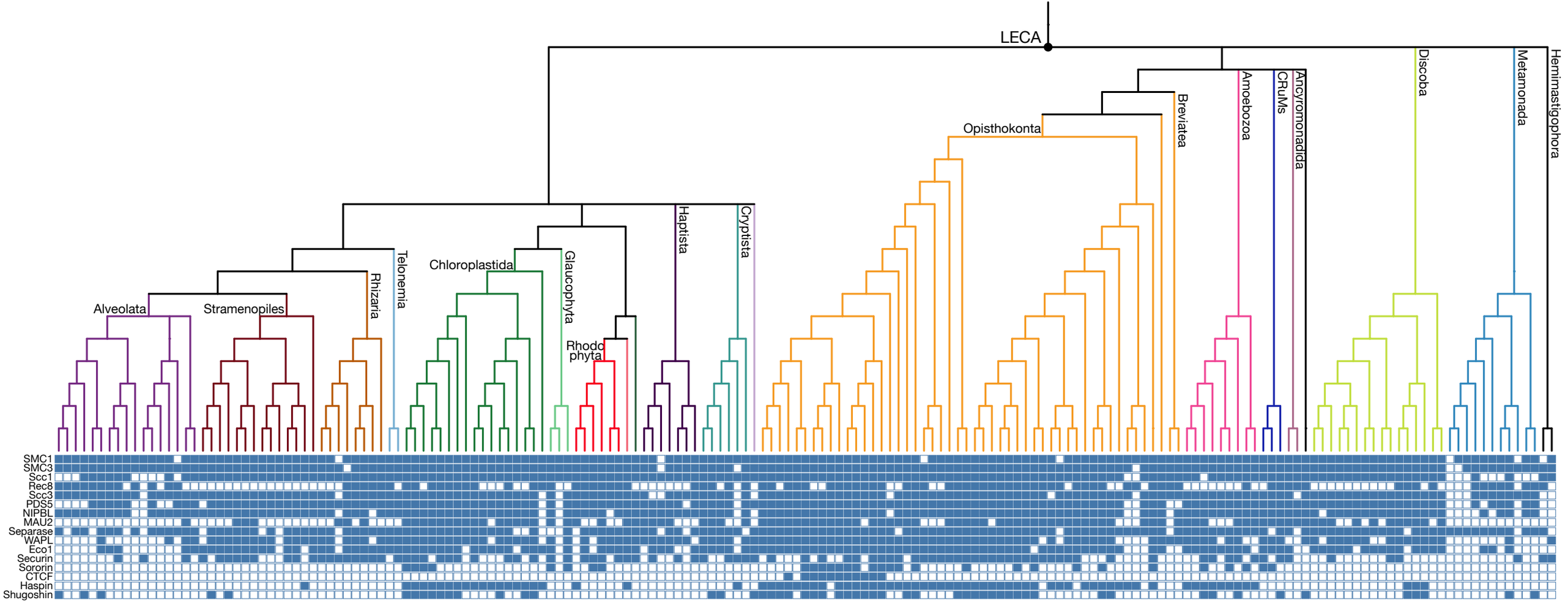
81. R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

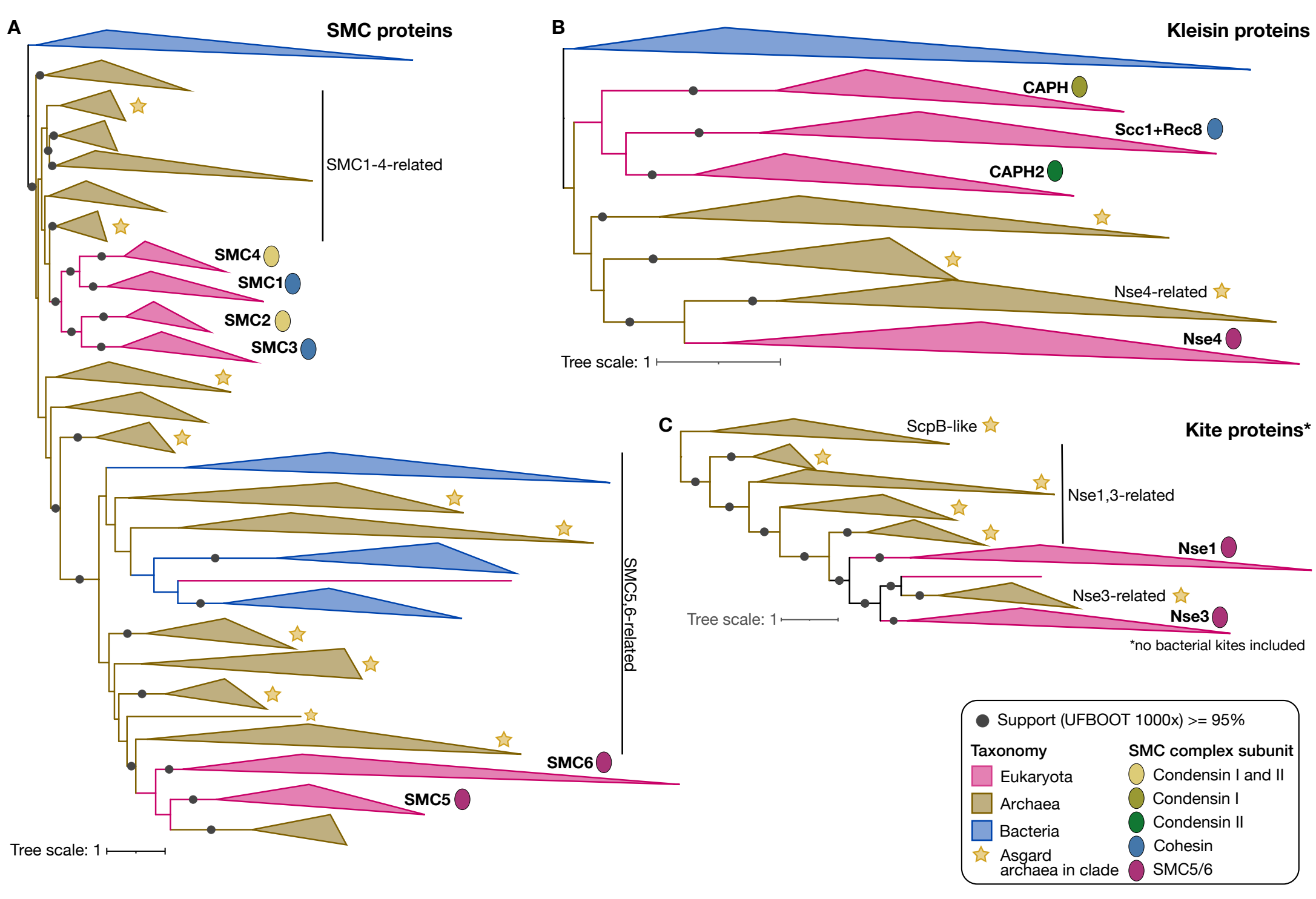

SMC complex composition
hinge
ν-SMC
κ-SMC
Kleisin
head
Hawks (condensin, cohesin)
Kites (SMC5/6, prokaryotes)
Condensin I
Condensin II
Cohesin
SMC5/6
1 Eukaryota
2 LECA
3
Asgard archaea
Bacteria
Archaea
LECA: Last Eukaryotic Common Ancestor

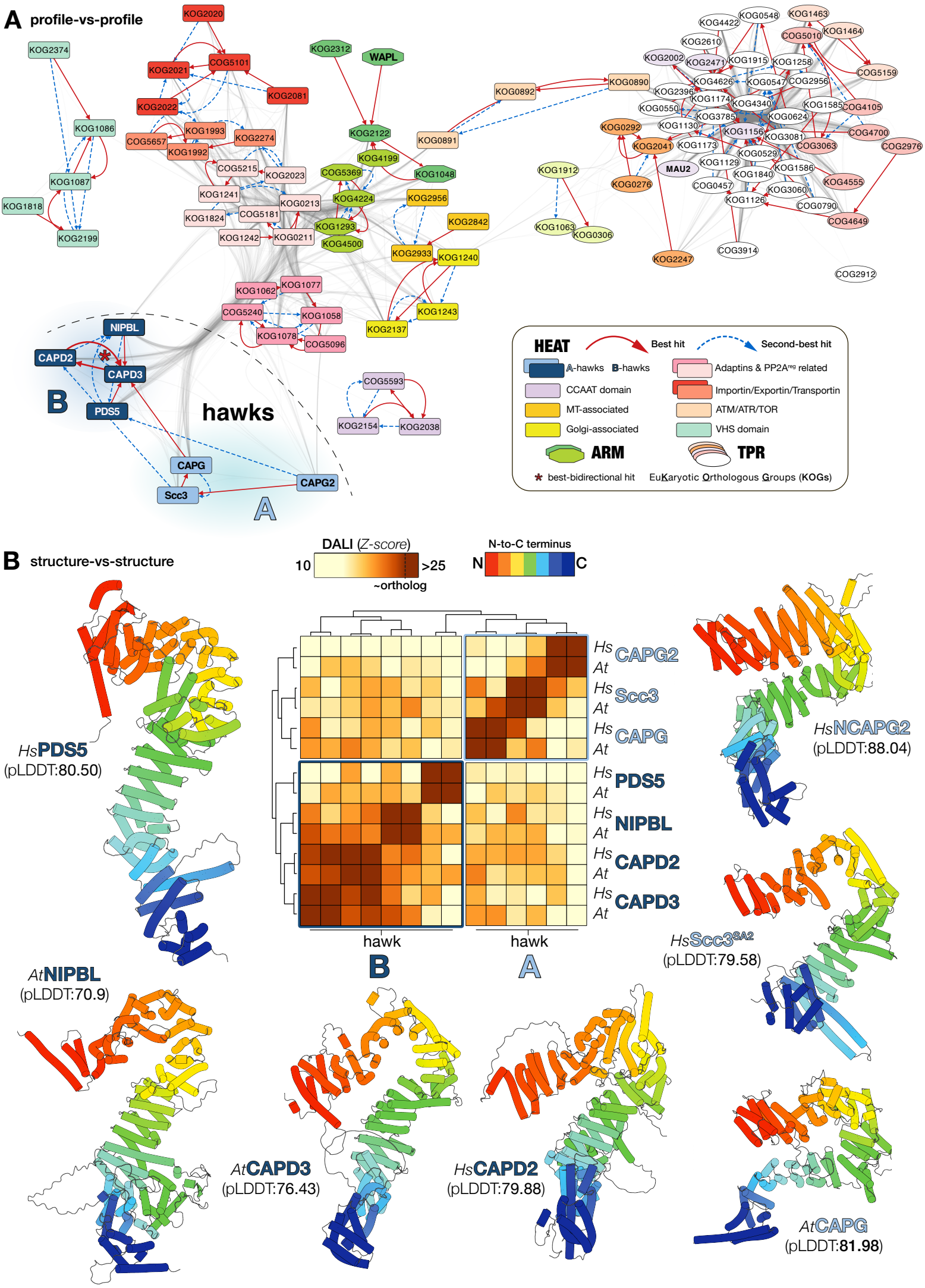


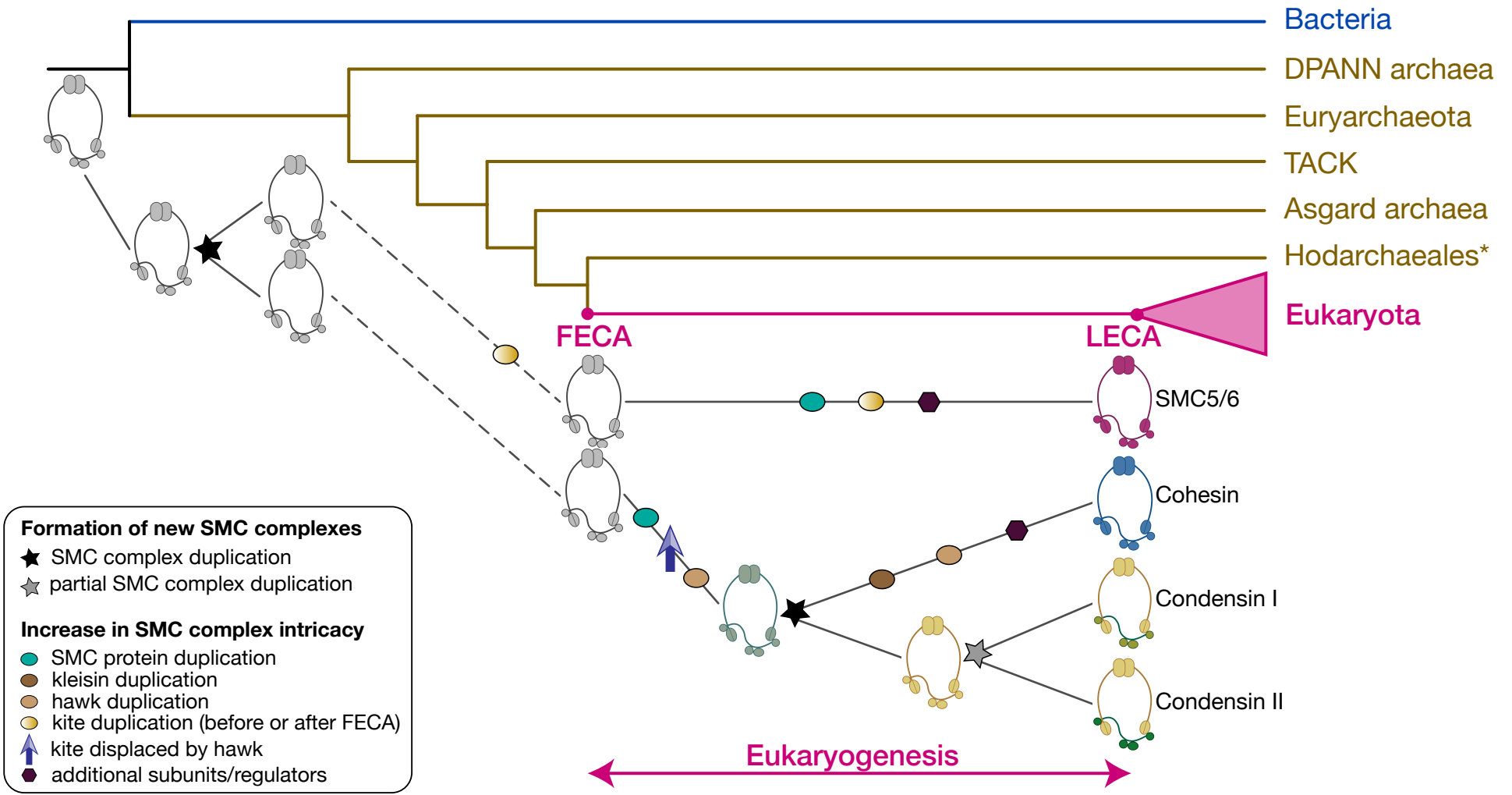
*Gonapodya prolifera (a fungus)


LECA
Alveolata
Stramenopiles
Rhizaria
Telonemia
Chloroplastida
Glaucophyta
Rhodo
phyta
Haptista
Cryptista
Opisthokonta
Breviatea
Amoebozoa
CRuMs
Ancyromonadida
Discoba
Metamonada
Hemimastigophora
SMC5
SMC6
Nse4
Nse1
Nse3
Nse2
Nse5
Nse6

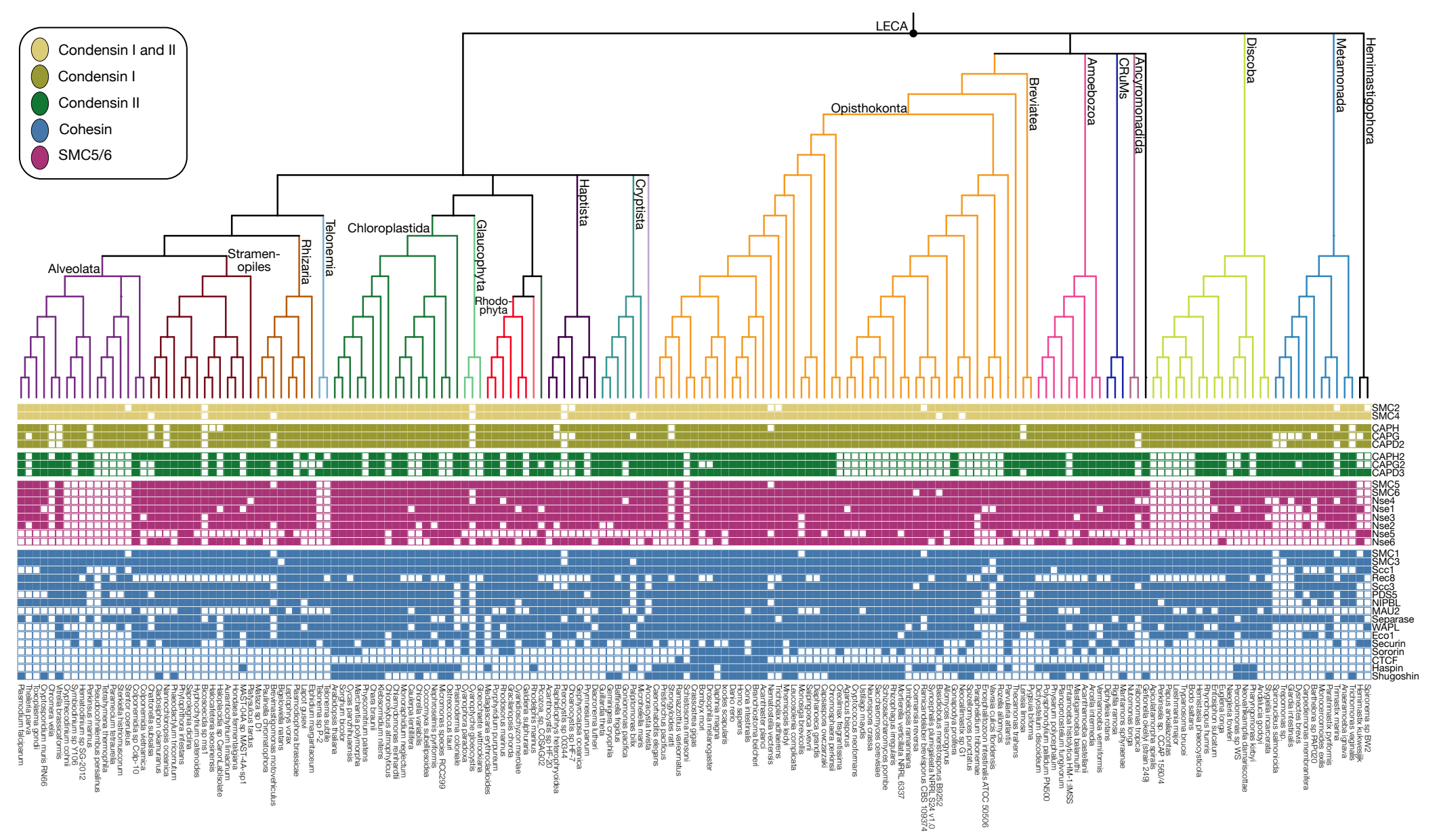

LECA
Alveolata
Stramenopiles
Rhizaria
Telonemia
Chloroplastida
Glaucophyta
Rhodophyta
Haptista
Cryptista
Opisthokonta
Breviatea
Amoebozoa
CRuMs
Ancyromonadida
Discoba
Metamonada
Hemimastigophora
SMC1
SMC3
Scc1
Rec8
Scc3
PDS5
NIPBL
MAU2
Separase
WAPL
Eco1
Securin
Sororin
CTCF
Haspin
Shugoshin

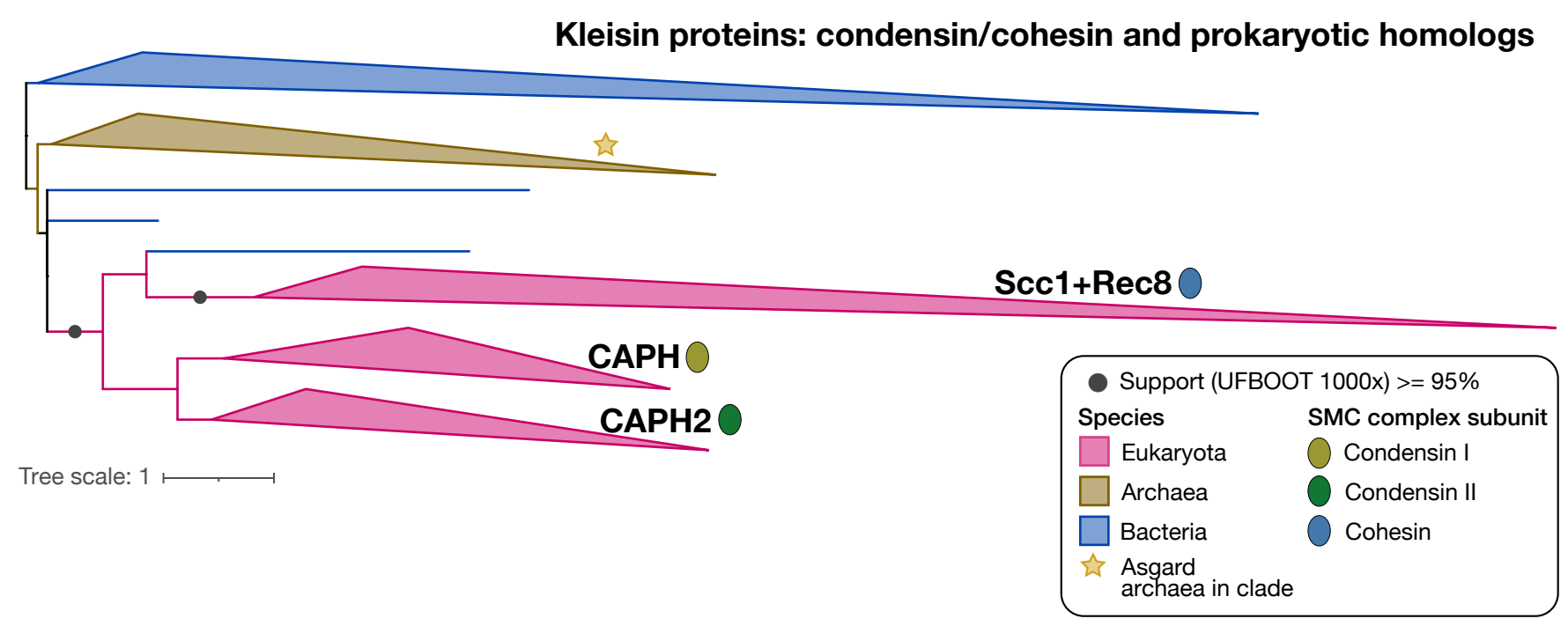

A
SMC proteins
SMC1-4-related
SMC4
SMC1
SMC2
SMC3
SMC5,6-related
SMC6
SMC5
Tree scale: 1
B
Kleisin proteins
CAPH
Scc1+Rec8
CAPH2
Nse4-related
Nse4
Tree scale: 1
C
Kite proteins*
ScpB-like
Nse1,3-related
Nse1
Nse3-related
Nse3
Tree scale: 1
*no bacterial kites included
Support (UFBOOT 1000x) >= 95%
Taxonomy
Eukaryota
Archaea
Bacteria
Asgard archaea in clade
SMC complex subunit
Condensin I and II
Condensin I
Condensin II
Cohesin
SMC5/6

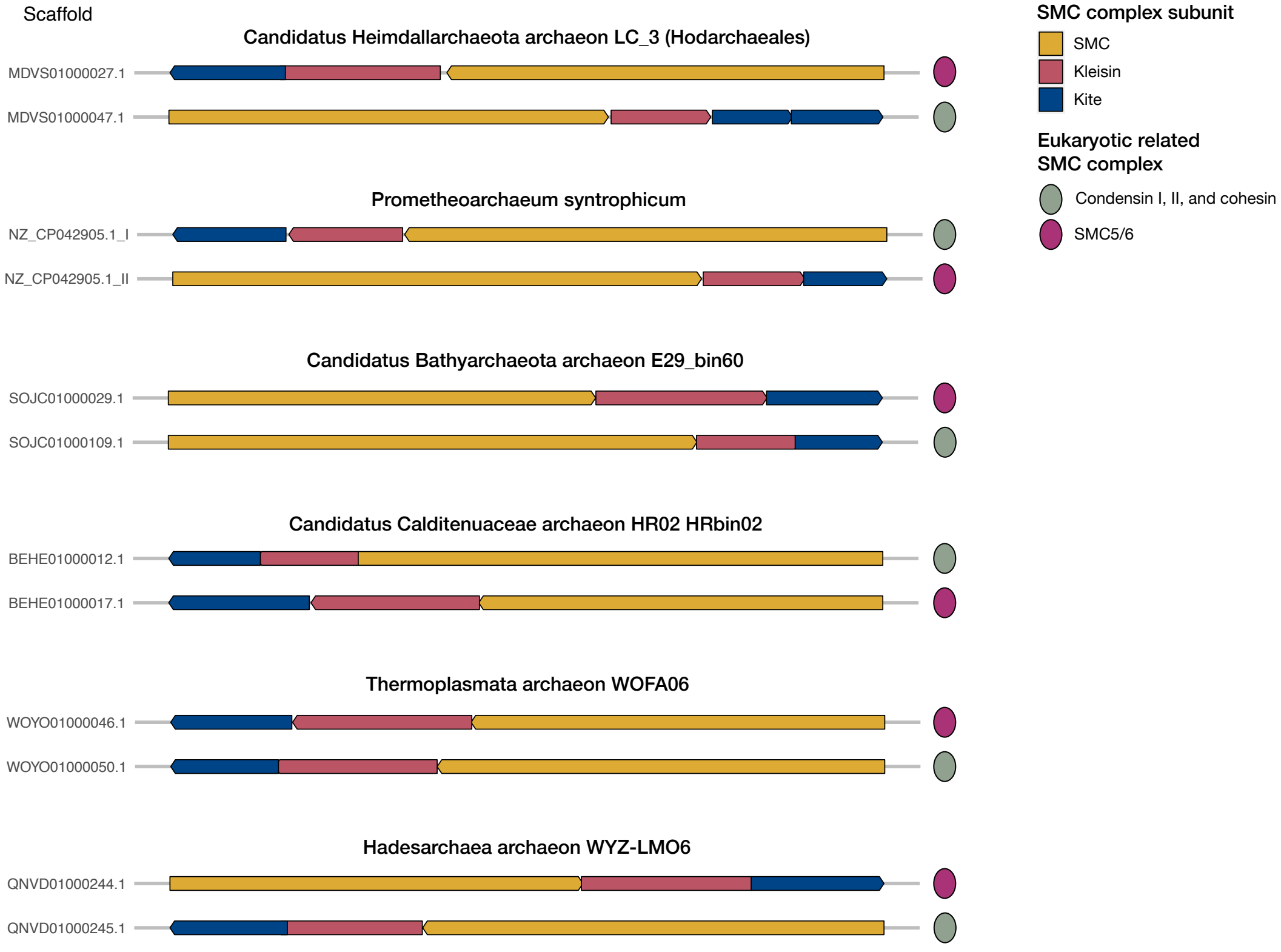

A profile-vs-profile
KOG2020
KOG2374
KOG2021
COG5101
KOG2081
KOG2022
KOG1086
KOG1993
COG5657
KOG2274
KOG1992
KOG1087
KOG1818
KOG2199
KOG2312
WAPL
KOG2122
KOG4199
COG5369
KOG1048
KOG0892
KOG0891
KOG0890
KOG1912
KOG0292
KOG2041
KOG0276
KOG1063
KOG0306
KOG2247
COG5215
KOG2023
KOG1241
KOG0213
KOG4224
KOG2956
KOG1824
COG5181
KOG1293
KOG2842
KOG1242
KOG0211
KOG4500
KOG2933
KOG1240
KOG1062
KOG1077
COG5240
KOG1058
KOG1243
KOG1078
COG5096
KOG2137
COG5593
KOG2154
KOG2038
NIPBL
CAPD2
CAPD3
PDS5
B
hawks
CAPG
CAPG2
Scc3
A
KOG4422
KOG0548
KOG1463
KOG2610
COG5010
KOG1464
KOG2002
KOG2471
KOG1915
KOG1258
COG5159
KOG4626
KOG0547
COG2956
KOG2396
KOG1174
KOG4340
KOG1585
COG4105
KOG0550
KOG3785
KOG0624
KOG1130
KOG1156
KOG3081
COG4700
KOG1173
COG3063
COG2976
KOG0529
MAU2
KOG1129
KOG1586
KOG1840
KOG4555
COG0457
KOG3060
KOG1126
COG0790
COG4649
COG3914
COG2912
HEAT
Best hit
Second-best hit
A-hawks
B-hawks
Adaptins & PP2A^reg related
CCAAT domain
Importin/Exportin/Transportin
MT-associated
ATM/ATR/TOR
Golgi-associated
VHS domain
ARM
TPR
best-bidirectional hit
EuKaryotic Orthologous Groups (KOGs)
B structure-vs-structure
DALI (Z-score)
10
>25
~ortholog
N-to-C terminus
N
C
Hs CAPG2
At
Hs Scc3
At
Hs CAPG
At
Hs PDS5
At
Hs NIPBL
At
Hs CAPD2
At
Hs CAPD3
At
hawk
B
hawk
A
HsPDS5
(pLDDT:80.50)
HsNCAPG2
(pLDDT:88.04)
HsScc3SA2
(pLDDT:79.58)
AtNIPBL
(pLDDT:70.9)
AtCAPD3
(pLDDT:76.43)
HsCAPD2
(pLDDT:79.88)
AtCAPG
(pLDDT:81.98)

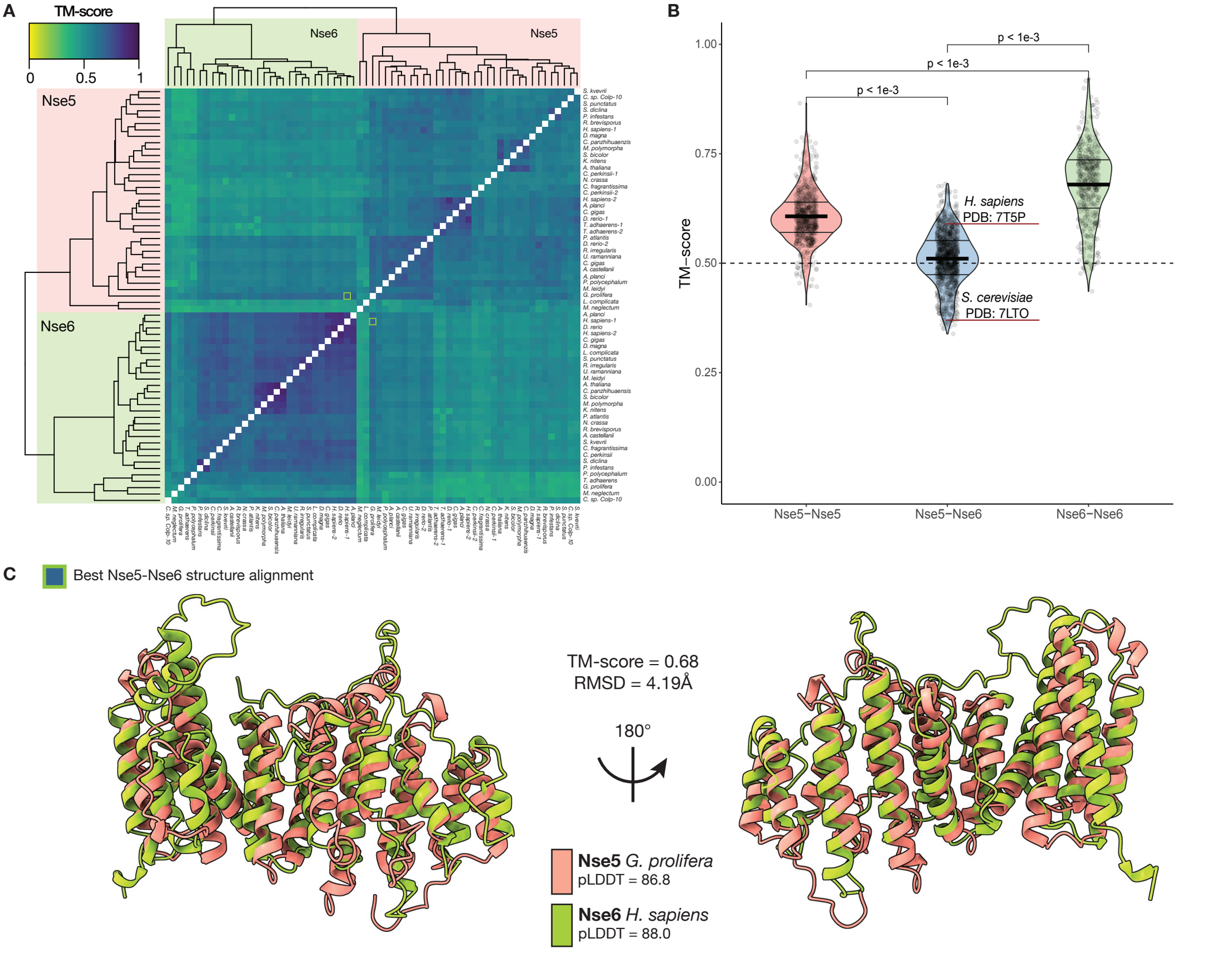


*Hodarchaeales: order within Asgard archaea, closest identified prokaryotic relatives of eukaryotes[6]


Condensin I and II
Condensin I
Condensin II
Cohesin
SMC5/6
LECA
Alveolata
Stramen-
opiles
Rhizaria
Telonemia
Chloroplastida
Glaucophyta
Rhodo-
phyta
Haptista
Cryptista
Opisthokonta
Breviatea
Amoebozoa
CRuMs
Ancyromonadida
Discoba
Metamonada
Hemimastigophora
SMC2
SMC4
CAPH
CAPG
CAPD2
CAPH2
CAPG2
CAPD3
SMC5
SMC6
Nse4
Nse1
Nse3
Nse2
Nse5
Nse6
SMC1
SMC3
Scc1
Rec8
Scc3
PDS5
NIPBL
MAU2
Separase
WAPL
Eco1
Securin
Sororin
CTCF
Haspin
Shugoshin
Sarcomonas sp BW2
Hemimastix kukwesjijk
Trichomonas vaginalis
Anaeramoeba ignava
Paratrimastix pyriformis
Barthelona sp PAP020
Monocercomonoides exilis
Carpediemonas membranifera
Giardia intestinalis
Trepomonas sp.
Spironucleus salmonicida
Shygella incarcerata
Hubudiu goodayi
Propavlovskella kirbyi
Neovahlkampfia damariscottae
Percolomonas sp WS
Naegleria fowleri
Euglena longa
Euglena gracilis
Rhynchopus humris
Hemistasia phaeocystis
Trypanosoma brucei
Papus ankistrodesmus
Bodo saltans
Perkinsela sp. CCAP 1560/4
Angomonas deanei
Geforella okelly (strain 249)
Angomonas sp. CCAP 1560/4
Nutomonas longa
Mertamonas sphyraenae
Diphylleia rotans
Rigifila ramosa
Acrolella intermedia
Acanthamoeba castellanii
Mastigamoeba balamuthi
Entamoeba histolytica HM-1:IMSS
Entamoeba dispar SAW760
Physarum polycephalum
Polysphondylium pallidum PN500
Dictyostelium discoideum
Pygsuia biforma
Thecamonas trahens
Pavlovalea atlantis
Rozella allomycis
Vavraia culicis floridensis
Encephalitozoon intestinalis ATCC 50506
Sacchromyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Ustilagomyces sp. PMS-2011
Cryptococcus neoformans
Aspergillus fumigatus
Coccidioides posadasii
Chromosphaera perkinsii
Creolimax fragrantissima
Capsaspora owczarzaki
Salpingoeca kvevri
Monosiga brevicollis
Mikrocystis aeruginosa
Leucosolenia complicata
Trichoplax adhaerens
Nematostella vectensis
Acanthaster planci
Branchiostoma belcheri
Ciona intestinalis
Danio rerio
Mus musculus
Homo sapiens
Ixodes scapularis
Daphnia magna
Drosophila melanogaster
Caenorhabditis elegans
Schistosoma mansoni
Strongyloides ratti
Ramazzottius varieornatus
Crassostrea gigas
Ancorayata twista
Microhelella maris
Palpitomonas bilix
Guillardia theta
Bellinella frigidus
Gemmigera cryophila
Deconema luthen
Euglenaria anomala
Gephyrocapsa oceanica
Chaetoceros sp. CCMP2037
Phaeodactylum tricornutum
Thalassiosira pseudonana
Nannochloropsis oceanica
Phytophthora infestans
Hyphochytrium catenoides
Aplanochytrium stocchinoi
Cladosiphon okamuranus
Ectocarpus siliculosus
Saccharina japonica
Nannochloropsis gaditana
Cyanophora paradoxa
Porphyridium purpureum
Rhodosorus marinus
Cyanidioschyzon merolae
Galdieria sulphuraria
Rhodelphis marinus
Cyanophora paradoxa
Gloeochaete witrockiana
Mesostigma viridis
Chlamydomonas reinhardtii
Volvox carteri
Monoraphidium neglectum
Chlorella variabilis
Coccomyxa subellipsoidea
Cauleria lentillifera
Ostreococcus lucimarinus
Micromonas pusilla
Micromonas sp. RCC299
Nephroselmis pyriformis
Pycnococcus provasoli
Ostreococcus tauri
Prasinoderma coloniale
Cyanophyceae glaucocystis
Cyanophyceae gloeocystis
Porphyridium purpureum
Rhodosorus marinus
Bracteacoccus minor
Chlorella variabilis
Chlorella sorokiniana
Chlorella vulgaris
Chlorella protothecoides
Chlorella sp. NC64A
Chlorella sp. ATCC 50258
Chlorella sp. ATCC 50258
Chlorella sp. ATCC 50258


Kleisin proteins: condensin/cohesin and prokaryotic homologs
Scc1+Rec8
CAPH
CAPH2
Tree scale: 1
Support (UFBOOT 1000x) >= 95%
Species
Eukaryota
Archaea
Bacteria
Asgard archaea in clade
SMC complex subunit
Condensin I
Condensin II
Cohesin


Scaffold
Candidatus Heimdallarchaeota archaeon LC_3 (Hodarchaeales)
MDVS01000027.1
MDVS01000047.1
Prometheoarchaeum syntrophicum
NZ_CP042905.1_I
NZ_CP042905.1_II
Candidatus Bathyarchaeota archaeon E29_bin60
SOJC01000029.1
SOJC01000109.1
Candidatus Calditenuaceae archaeon HR02 HRbin02
BEHE01000012.1
BEHE01000017.1
Thermoplasmata archaeon WOFA06
WOYO01000046.1
WOYO01000050.1
Hadesarchaea archaeon WYZ-LMO6
QNVD01000244.1
QNVD01000245.1
SMC complex subunit
SMC
Kleisin
Kite
Eukaryotic related SMC complex
Condensin I, II, and cohesin
SMC5/6


A
TM-score
0
0.5
1
Nse6
Nse5
Nse5
Nse6
B
1.00
0.75
0.50
0.25
0.00
TM−score
p < 1e-3
p < 1e-3
p < 1e-3
H. sapiens
PDB: 7T5P
S. cerevisiae
PDB: 7LTO
Nse5−Nse5
Nse5−Nse6
Nse6−Nse6
C
Best Nse5-Nse6 structure alignment
TM-score = 0.68
RMSD = 4.19Å
180°
Nse5 G. prolifera
pLDDT = 86.8
Nse6 H. sapiens
pLDDT = 88.0