

1 **The first eukaryotic kinome tree illuminates the dynamic history of**
2 **present-day kinases**

3

4 Leny M. van Wijk¹, Berend Snel^{1*}

5

6 ¹Theoretical Biology and Bioinformatics, Department of Biology, Science Faculty, Utrecht

7 University, Utrecht, The Netherlands. *email: b.snel@uu.nl

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25 **Abstract**

26 Eukaryotic Protein Kinases (ePKs) are essential for eukaryotic cell signalling. Several phylogenetic
27 trees of the ePK repertoire of single eukaryotes have been published, including the human kinome
28 tree. However, a eukaryote-wide kinome tree was missing due to the large number of kinases in
29 eukaryotes. Using a pipeline that overcomes this problem, we present here the first eukaryotic
30 kinome tree. The tree reveals that the Last Eukaryotic Common Ancestor (LECA) possessed at least
31 92 ePKs, much more than previously thought. The retention of these LECA ePKs in present-day
32 species is highly variable. Fourteen human kinases with unresolved placement in the human
33 kinome tree were found to originate from three known ePK superfamilies. Further analysis of ePK
34 superfamilies shows that they exhibit markedly diverse evolutionary dynamics between the LECA
35 and present-day eukaryotes. The eukaryotic kinome tree thus unveils the evolutionary history of
36 ePKs, but the tree also enables the transfer of functional information between related kinases.

37

38 **Introduction**

39 Kinases are fundamental to convey information in living organisms. Due to their importance for
40 health and agriculture, they are studied in a wide variety of eukaryotic species¹⁻⁶. In eukaryotes,
41 the vast majority of kinases belongs to a single family: the eukaryotic Protein Kinases (ePKs)^{7,8}.
42 EPKs are characterised by a conserved kinase domain of about 250 amino acids that consists of 12
43 subdomains⁹. They phosphorylate either serine/threonine or tyrosine residues or have dual
44 specificity. EPKs are subdivided into seven superfamilies: AGC, CAMK, CK1, CMGC, STE, Tyrosine
45 Kinase (TK) and Tyrosine Kinase-Like (TKL)^{7,9}. EPKs that lack a superfamily have previously been
46 referred to as 'Other'⁷, while in this paper they will be referred to as Unaffiliated.

47

48 In 2002, a paper on the protein kinase complement of the human genome was published⁷. This
49 highly cited paper was accompanied by an iconic poster with a phylogenetic tree of all human ePK
50 domains. Kinome analyses of several other eukaryotic species, sometimes including a species-
51 specific kinome tree, followed¹⁰⁻¹⁴. Currently, genomic data is available for a broad diversity of
52 eukaryotic species. However, available kinome trees never incorporated how kinases are related
53 across the entire eukaryotic tree of life. This is unfortunate as a eukaryotic kinome tree is relevant
54 both to understand the function of particular ePKs better and to reveal the ancient evolutionary
55 history of ePKs.

56

57 The functional relevance of a eukaryotic kinome tree lies foremost in facilitating the transfer of
58 functional information between neighbouring proteins. Proteins that are close to each other in a
59 phylogenetic tree potentially share conserved molecular interactions and mechanical properties
60 that arose in their common ancestor. Therefore, information about well-studied ePKs in a
61 eukaryotic kinome tree can be used to generate hypotheses for the study of related ePKs. Even
62 ePKs that are not included in a eukaryotic kinome tree can benefit from the transfer of functional
63 information: proteins that form a single branch in a phylogenetic tree enable classifying proteins
64 outside the tree using Hidden Markov Models (HMMs).

65

66 A eukaryotic kinome tree is also important for the evolutionary cell biology of eukaryotes¹⁵. For
67 example, current estimates of the number and identity of kinases that were already present in the
68 Last Eukaryotic Common Ancestor (LECA) are only based on limited sets of eukaryotic species^{7,16}. A
69 eukaryotic kinome tree allows to determine the ePK complement in the LECA more precisely.

70

71 Although a eukaryotic kinome tree is relevant both from a functional and an evolutionary

72 perspective, there is one substantial hurdle: the large number of kinases in eukaryotes. Only the
73 human kinome tree consists already of 491 ePK domains⁷, and in a collection of nearly 100
74 eukaryotes, this number increases to over 36,000 ePK domains. Such a number of sequences
75 precludes the use of state-of-the-art alignment as well as tree building software. Moreover, it is a
76 Sisyphean task to analyse a phylogenetic tree that consists of over 36,000 leaves.

77

78 A more general problem of gene trees is the negative impact of rapidly evolving sequences on
79 statistic support. A commonly used strategy to improve statistic support in species trees is to
80 select slowly evolving genes or positions¹⁷. For gene trees, an equivalent of this strategy has been
81 proposed: the ScrollSaw method¹⁸. The ScrollSaw method systematically selects only slowly
82 evolving sequences for generating a gene tree. As a result, both the number of sequences is
83 reduced, and rapidly evolving sequences are excluded. This makes the ScrollSaw method perfectly
84 suitable to handle the large number of ePKs and generate a well-supported eukaryotic kinome
85 tree.

86

87 Here we present the first eukaryotic kinome tree, generated with a modified and extended version
88 of the ScrollSaw method. The tree reveals ePK superfamily membership for several ePKs that have
89 been labelled as Unaffiliated in the human kinome tree, most notably CAMKK1 and CAMKK2. The
90 tree furthermore unveils that the LECA had much more ePKs than was thought before: at least 92.
91 These 92 ePKs include some surprising examples of ePKs that were previously believed to be
92 specific for certain eukaryotic supergroups, like human CHK1 and the plant CIPKs. The number of
93 LECA ePKs retained in present-day species varies enormously within and between eukaryotic
94 supergroups. The expansion of LECA ePKs since the common ancestor of eukaryotes also differs
95 within and between ePK superfamilies. This variation in LECA ePK dispensability and duplicability is

96 possibly linked to differential roles of LECA ePKs in cellular housekeeping and organismal
97 innovation. The eukaryotic kinome tree thus both reveals the evolutionary history of ePKs and
98 directs the study of ePK function.

99

100 **Results**

101 **The eukaryotic kinome tree reveals well-supported LECA kinase clades**

102 In order to generate a eukaryotic kinome tree, 36,475 ePK domains were collected from 94
103 eukaryotic species (Supplementary Data 1). These domains were used as input for a pipeline that
104 generates two phylogenetic trees by implementing a modified and extended version of the
105 Scrollsaw method (Fig. 1, Methods). In this LECA clade annotation pipeline, the Scrollsaw method
106 was extended with automatic annotation of the tree leaves into LECA kinase clades: groups of
107 kinases that likely have a single ancestor in the LECA because they include at least one Amorphea
108 and one Bikonta sequence¹⁹. The LECA clade annotation pipeline generated two different
109 eukaryotic kinome trees in order to facilitate automatic annotation, cross-validate annotated LECA
110 kinase clades and produce HMM profiles that contain more sequence diversity. One tree is based
111 on Bi-directional Best Hits (BBHs) between two eukaryotic supergroups (Fig. 2, Supplementary
112 Data 2-5), while the other tree is based on BBHs between five eukaryotic supergroups
113 (Supplementary Fig. 1, Supplementary Data 6-9). The leaves of both trees were automatically
114 annotated into LECA kinase clades, and both sets of LECA kinase clades were combined into one
115 non-overlapping set. This combined set was improved manually, resulting in a final set of 118 LECA
116 kinase clades.

117

118 Even though the LECA lived about 1-1.9 billion years ago²⁰, the vast majority of LECA kinase clades
119 in the eukaryotic kinome tree is statistically well supported. For example, in the two-supergroups-

120 BBHs tree, 78 per cent of the LECA kinase clades have bootstrap support values of 95 or above
121 (Fig. 2). This percentage is lower for the five-supergroups-BBHs tree (Supplementary Fig. 1). Higher
122 support for LECA kinase clades in the two-supergroups-BBHs tree is in agreement with results from
123 the original Scrollsaw paper¹⁸. There, bootstrap support for LECA clades increased upon additional
124 reduction of BBHs to one sequence per eukaryotic supergroup per LECA clade. The high bootstrap
125 support for our LECA kinase clades confirmed the usefulness of the Scrollsaw method to obtain
126 well-supported orthologous clades in large gene families.

127

128 Although LECA kinase clades are well supported, internal support in the eukaryotic kinome tree is
129 lower (Supplementary Results). Bootstrap support values of at least 70 are found only in 26 per
130 cent of the 112 pre-LECA duplications in the two-supergroups-BBHs tree (Fig. 2). Surprisingly, 34
131 per cent of these well-supported pre-LECA kinase clades appear deeper in the tree and include
132 three or more LECA kinase clades.

133

134 **Nearly 80 per cent of kinase domains can be assigned to a LECA kinase clade**

135 Protein assignment via HMM profiles outperforms assignment via BLAST²¹. Therefore, HMM
136 profiles of the 118 LECA kinase clades of the eukaryotic kinome tree (Supplementary Data 10)
137 were used to automatically assign the initial set of 36,475 kinase domains (Fig. 1, Supplementary
138 Table 1, Supplementary Data 11). Despite a conservative approach, 28,893 kinase domains were
139 assigned to their top hitting LECA kinase clade. Kinase domains were not assigned if the top two
140 scoring LECA kinase clades had a bit score difference below 10 or a maximal bit score below 30
141 (Supplementary Data 12 and 13). Although the assigned kinase domains encompass nearly 80 per
142 cent of the total kinase domain dataset, there is considerable variation in assignment percentages
143 between species (Supplementary Table 2). For each LECA kinase clade, the protein name of the

144 best scoring human kinase domain was used to name the LECA kinase clade. If human hits were
145 not available, best hits from baker's yeast or *Arabidopsis thaliana* were used for naming. LECA
146 kinase clades to which no kinase domains from these three species were assigned are indicated
147 with Orthologous Group (OG) and a number.

148

149 **LECA complexity involved at least 92 eukaryotic protein kinases**

150 Our reconstruction of LECA kinase clades enabled estimating the LECA kinase repertoire. However,
151 not all 118 LECA kinase clades are equally likely to represent a single gene in the ancestor of all
152 eukaryotes. An initial conservative estimate of the most reliable LECA kinase clades yielded 91
153 ePKs in the LECA. These 91 LECA kinases correspond to 91 LECA kinase clades that are annotated
154 in both eukaryotic kinome trees (Fig. 2, Supplementary Fig. 1), have at least one bootstrap support
155 of minimal 70, and kinase domains from minimal two eukaryotic supergroups have been assigned
156 to them (Supplementary Data 14). Other LECA kinase clades did not fulfil one or more of these
157 criteria and require more investigation. For example, LECA kinase clades to which only a limited
158 number of kinase domains were assigned need closer examination. This could discriminate
159 between possible explanations like horizontal gene transfer, genome contamination, or being a
160 bonafide LECA kinase that has been lost in many species.

161

162 In addition to the 91 LECA ePKs that are based on LECA kinase clades, one more LECA kinase was
163 added: Haspin. This kinase was absent in our kinase domain dataset because it is a diverged ePK
164 with a PFAM model distinct from the PFAM models that were used to collect sequences for the
165 eukaryotic kinome tree⁸. Haspin was likely already present in the LECA^{16,22}. Thus our initial
166 conservative estimate of 91 ePKs in the LECA together with Haspin result in an estimate of 92 ePKs
167 in the LECA. This is more than a third more than the largest previous estimate of 68 basal

168 ePKs¹⁶ (Supplementary Results). A large LECA ePK complement is in line with a LECA that was much
169 more complex than many present-day eukaryotes²⁰.

170
171 **The common ancestry of LKB1 and CAMKKs explains their functional overlap**

172 The eukaryotic kinome tree is highly consistent with the human kinome tree, but a complete
173 agreement would require some adjustments to the human kinome tree (Supplementary Results).
174 The eukaryotic kinome tree, for example, clarified the relationships between a few Unaffiliated
175 human kinases and the ePK superfamilies (Supplementary Results). The most interesting example
176 of Unaffiliated human kinases that stem from within an ePK superfamily are the human kinases
177 assigned to LECA kinase clade CAMKK2 (Fig. 2, Supplementary Fig. 1). The names of these kinases,
178 CAMKK1 and CAMKK2, already reflect their functional link with the CAMK superfamily. Despite this
179 functional link, CAMKK1 and CAMKK2 were placed at a distance from the CAMK superfamily in the
180 human kinome tree (Supplementary Fig. 2). In contrast, LECA kinase clade CAMKK2 is positioned
181 within the CAMK superfamily in the eukaryotic kinome tree. It is located next to LECA kinase clade
182 LKB1 with high bootstrap support (99; Fig. 2).

183
184 The juxtaposition of LECA kinase clades LKB1 and CAMKK2 is striking both from a functional and
185 evolutionary perspective. In human, LKB1 is a master kinase of AMPK and the AMPK-related
186 kinases²³, which were assigned to the LECA kinase clades AMPKA2, MARK2 and BRSK1
187 (Supplementary Table 1). In addition to LKB1, AMPK can also be phosphorylated by CAMKK2²⁴. In
188 baker's yeast and *A. thaliana*, orthologs of CAMKK1 and CAMKK2 are the only upstream kinases of
189 AMPK orthologs because LKB1 is absent in these species²⁵ (Supplementary Table 1). The
190 juxtaposition of LECA kinase clades LKB1 and CAMKK2 is therefore in agreement with their

191 overlapping function in AMPK phosphorylation. It suggests that the common ancestor of LECA
192 kinase clades LKB1 and CAMKK2 may already have been able to phosphorylate the common
193 ancestor of LECA kinase clades AMPKA2, MARK2 and BRSK1. The common ancestor of LKB1,
194 CAMKK2 and possibly OG040 may even have been able to phosphorylate the common ancestor of
195 all other CAMK LECA kinase clades. This is suggested by the basal position of LKB1, CAMKK2 and
196 OG040 within the CAMK superfamily (Fig. 2, Supplementary Fig. 1), and the fact that CAMKK2 can
197 also phosphorylate members of LECA kinase clade CAMK1D²⁶. Interestingly, also within the AGC
198 superfamily, the most basal position is reserved for a master kinase: PDK1²⁷ (Fig. 2, Supplementary
199 Fig. 1).

200

201 **The majority of Unaffiliated human kinases stem from Unaffiliated kinase clades**

202 One of our reasons to generate the eukaryotic kinome tree was to test whether there exist
203 Unaffiliated human kinases that actually belong to an ePK superfamily. Including CAMKK1 and
204 CAMKK2, 14 of the 88 Unaffiliated human kinases could be classified into a superfamily
205 (Supplementary Results). However, no less than 56 Unaffiliated human kinases were assigned to
206 17 Unaffiliated (pre-)LECA kinase clades that have no well-supported further affiliations in the
207 eukaryotic kinome tree (Supplementary Results). Although their phylogenetic position might be
208 insufficiently resolved due to accelerated evolution, many of the Unaffiliated (pre-)LECA kinase
209 clades could also be the result of old duplications early in eukaryogenesis. Such an old age could
210 explain why it is difficult to connect Unaffiliated (pre-)LECA kinase clades firmly to any other
211 (pre-)LECA kinase clade or ePK superfamily. The Unaffiliated (pre-)LECA kinase clades are then as
212 old and distinct as entire ePK superfamilies but duplicated less vigorously during eukaryogenesis
213 than most ePK superfamilies.

214

215 **Human CHK1 and plant CIPKs were one kinase in the LECA**

216 The LECA kinase clade delineation suggested a single LECA ancestor for kinases from different
217 eukaryotic groups that so far were thought to be group-specific. A prominent case is the common
218 ancestry of plant CBL-Interacting Protein Kinases (CIPKs) and opisthokont Checkpoint Kinase 1
219 (CHK1). CIPKs, also known as SNRK3s, have often been described as plant-specific, but recently
220 they have also been found in other eukaryotic species²⁸. CHK1 is considered opisthokont-specific²⁹.
221 However, both kinase families were assigned to LECA kinase clade CHK1 within the CAMK
222 superfamily, suggesting a common ancestor in the LECA (Supplementary Table 1, Fig. 2,
223 Supplementary Fig. 1). This suggestion is supported by some additional experimental and
224 phylogenetic evidence^{21,29-33}.

225

226 The CIPK-specific NAF domain is absent in opisthokont CHK1. However, this domain is still present
227 in *Thecamonas trahens* CHK1 (Supplementary Data 15). *T. trahens* is a member of the Apusozoa,
228 the sister clade of opisthokonts, and therefore belongs to the Amorphea. The presence of the NAF
229 domain in both Amorphea CHK1 and Bikonta CIPKs suggests that the LECA CHK1 had a NAF domain
230 as well. The absence of the NAF domain in opisthokont CHK1 is, therefore, a derived feature due
231 to a loss in the common ancestor of fungi and animals.

232

233 **Human HIPKs and baker's yeast YAK1 were one kinase in the LECA**

234 Another unrecognised deep orthology was found between the human Homeodomain-Interacting
235 Protein Kinases (HIPKs) and baker's yeast Yet Another Kinase 1 (YAK1). YAK1 orthologs are present
236 throughout eukaryotes, but they were thought to be missing in Metazoa¹⁶. In KinBase³⁴, the HIPK
237 and YAK subfamilies have a perfectly complementary distribution: either Metazoa-specific or
238 missing in Metazoa. However, both human HIPKs and baker's yeast YAK1 were assigned to LECA

239 kinase clade HIPK2 within the CMGC superfamily (Supplementary Table 1, Fig. 2, Supplementary
240 Fig. 1).

241

242 In earlier studies, the HIPKs and YAK1 have been suggested to be different classes of DYRK
243 proteins^{35,36}. However, depending on how the phylogenetic trees in those studies are rooted, the
244 HIPKs and YAK1 could be inferred to be monophyletic. HIPK2 and YAK1 also have a shared function
245 in phosphorylating the CCR4-NOT complex³⁷. Together these data suggest that the metazoan HIPKs
246 and the eukaryote-wide found YAK1 were indeed one kinase in the LECA. The fact that metazoan
247 HIPKs apparently are difficult to recognise as YAK1 orthologs indicates derived characteristics for
248 HIPKs.

249

250 **The LECA kinase presence-absence profile displays diverse patterns of kinase retention**

251 The assignment of the initial set of 36,475 ePK domains to LECA kinase clades allowed the
252 generation of a clustered presence/absence profile of LECA kinases in present-day species (Fig. 3).
253 This clustering divided LECA kinases into two large clusters. The 'ubiquitous' cluster at the top of
254 the presence/absence profile contains 49 LECA kinases, of which 48 are present in at least half of
255 the 94 species. Four LECA kinases are retained in all extant eukaryotic species in our dataset: CDK3
256 (but note that CDK3 might comprise two nested LECA kinase clades, see Supplementary Results),
257 CAMK1D, CK2A1 and CK1D. Several other LECA kinases are nearly omnipresent: AURA, CRK7,
258 GSK3A, ERK5, MAP2K1, CAMKK2, SRPK1, PDK1, NDR2 and AMPKA2.

259

260 The cluster at the bottom of the presence/absence profile can be subdivided into two subclusters.
261 The first 'fungal-loss' subcluster contains 33 LECA kinases, of which 13 are present in at least half
262 of the 94 species, just like the kinases in the 'ubiquitous' cluster. However, many of the LECA

263 kinases in the 'fungal-loss' subcluster were lost in several or all fungi. The second 'sparse'
264 subcluster contains 36 LECA kinases, of which 30 are present in less than a quarter of the 94
265 species. Not surprisingly, the 'sparse' cluster encompasses the majority of LECA kinases that were
266 excluded from the LECA kinase number estimate because they were not present in a sufficient
267 number of species (indicated with **).

268

269 Within the 'sparse' subcluster, OG040 is particularly interesting. This LECA kinase is nearly only
270 found in early-branching species: two excavates, *Guillardia theta*, *Cyanophora paradoxa*, red
271 algae, two amoebae and *Capsaspora owczarzaki*. OG040 also has an interesting position in the
272 eukaryotic kinome tree: it clusters next to the master kinase clade that contains CAMKK2 and LKB1
273 (Fig. 2, Supplementary Fig. 1). The retention pattern and phylogenetic position of OG040 raise
274 curiosity about its function in the LECA and in present-day species.

275

276 **LECA kinase retention varies between and within eukaryotic supergroups**

277 The differential presence of eukaryotic supergroups in the 'ubiquitous', 'fungal-loss' and 'sparse'
278 cluster is reflected in Fig. 4, where species are ordered by their total number of LECA kinases.
279 Holozoa, a group that comprises animals and their unicellular relatives, dominate the top of this
280 graph. They are headed by *Branchiostoma floridae*, which shares the maximum number of 83
281 retained LECA kinases with the cryptophyte *G. theta*. Early-branching unicellular Holozoa are not
282 found among the top-scoring Holozoa, but the Choanomonadida *Salpingoeca rosetta* and
283 *Monosiga brevicollis* still kept 78 and 76 LECA kinases respectively. *C. owczarzaki* kept only 68
284 LECA kinases and lost quite some LECA kinases that are present in most Holozoa. However, it also
285 retained several LECA kinases that were lost in other Holozoa, like CDC15, FPK1 and IKS1.

286

287 The other large group within the opisthokonts, the fungi, exhibit a pattern strikingly different from
288 the Holozoa: they are mainly present at the bottom of Fig. 4. The fungi that kept most LECA
289 kinases are relatively early-branching species like the Chytridiomycota *Batrachochytrium*
290 *dendrobatidis* (67 LECA kinases) and *Spizellomyces punctatus* (66 LECA kinases). Within the
291 Archaeplastida, also early-branching species like the green alga *Chlamydomonas reinhardtii* and
292 charophyte alga *Klebsormidium flaccidum* kept the largest number of about 70 LECA kinases.
293 Interestingly, the model organisms *Saccharomyces cerevisiae* and *A. thaliana* maintained a
294 relatively small number of respectively 44 and 58 LECA kinases. At the very bottom of Fig. 4,
295 intracellular parasites with streamlined genomes like the fungal Microsporidia *Vavraia culicis* and
296 *Encephalitozoon intestinalis* are found³⁸. They kept less than 20 LECA kinases.

297

298 The small numbers of 39-43 LECA kinases that have been retained in the red algae *Chondrus*
299 *crispus*, *Galdieria sulphuraria*, *Cyanidioschyzon merolae* and *Porphyridium purpureum* can
300 probably also be attributed to a genome reduction³⁹. However, the red algae, just like *C.*
301 *owczarzaki* within the Holozoa, also reflect their early-branching position within the
302 Archaeplastida. Together with *C. paradoxa*, they kept three LECA kinases that were lost in the rest
303 of the Archaeplastida lineage. These three LECA kinases, LKB1, its neighbour OG040 and its
304 downstream kinase BRSK1, are all members of the CAMK superfamily. A fourth LECA kinase, PAK3
305 from the STE superfamily, is kept only in red algae but not in *C. paradoxa*. Interestingly, a human
306 kinase assigned to LECA kinase clade PAK3 is inhibited by LKB1⁴⁰. Therefore all four LECA kinases
307 that have been retained only in basal Archaeplastida are phylogenetically or functionally related.
308 This suggests that in these species, they may participate in the same process.

309

310 **The largest LECA kinase superfamily CMGC expanded least from LECA to human**

311 The eukaryotic kinome tree together with the presence-absence profile of its LECA kinases
312 enabled to quantify the evolutionary dynamics of ePK superfamilies from LECA till present-day
313 species. Except for the CMGC and CK1 superfamilies, ePK superfamilies had about 10 members in
314 the LECA (Fig. 2, Supplementary Fig. 1, Fig. 5). The CMGC superfamily was much larger with 24
315 LECA kinases while the CK1 superfamily was much smaller with two kinases in the LECA. Although
316 most kinase superfamilies had comparable sizes in the LECA, their expansion from LECA to human
317 is strikingly different (Fig. 5). The large CMGC superfamily expanded 2.8 times from LECA to
318 human. The small CK1 superfamily and medium-sized STE and AGC superfamilies are about five
319 times larger in human compared to the LECA. The other medium-sized superfamilies expanded
320 about 10 times (CAMK) or even more (TK/TKL).

321

322 **Kinase superfamily duplicability and dispensability are variable**

323 In general, superfamily expansion from LECA to human and the average expansion of single LECA
324 kinases from that same superfamily in 94 eukaryotes display a similar trend (Fig. 5, Fig. 6). For
325 example, LECA kinases from the large CMGC superfamily did not expand much from LECA to
326 human, and their per LECA kinase average expansion in 94 eukaryotes is also low. However, kinase
327 expansion from LECA to human and other present-day species is also variable within superfamilies.

328

329 The most striking variation in LECA kinase expansion is found within the TK/TKL and CAMK
330 superfamilies. These superfamilies expanded most from LECA to human (Fig. 5). However, the
331 average gene count of TK/TKL and CAMK LECA kinases in present-day eukaryotes is predominantly
332 low (<2) or high (>10) (Fig. 6). Some TK/TKL and CAMK LECA kinases, including TK/TKL LIMK1 and
333 CAMK AMPKA2, hardly expanded from LECA to present-day eukaryotes. These kinases are likely to
334 perform similar functions in extant eukaryotes as in the LECA. On the other hand, LECA kinases like

335 TK/TKL IRAK 4 and CAMK CAMK1D, underwent much duplication. Descendants of these LECA
336 kinases likely perform various new functions.

337

338 LECA kinases from the CMGC superfamily are special: they exhibit together with low duplicability
339 also low dispensability (Fig. 6). The average expansion of CMGC kinases from LECA to present-day
340 eukaryotes is just above 2, while nearly half of the CMGC LECA kinases are present in more than
341 80 present-day eukaryotes. LECA kinases from the TK/TKL superfamily display in contrast to the
342 CMGC superfamily both high duplicability and high dispensability (Fig. 6).

343

344 **Unaffiliated LECA kinases display low duplicability**

345 In the eukaryotic kinome tree, a large group of 53 LECA kinases is not part of any of the
346 superfamilies (Fig. 2, Supplementary Fig. 1, Fig. 5). A comparison with the human kinome tree
347 suggested that most of these Unaffiliated LECA kinases duplicated infrequently. The expansion of
348 the Unaffiliated LECA kinases from LECA to human is indeed very low, even below that of the
349 CMGC kinases (Fig. 5). However, the Unaffiliated LECA kinases and the CMGC superfamily display
350 different relationships between duplicability and dispensability. The Unaffiliated LECA kinases
351 generally combine low duplicability like the CMGC LECA kinases with high dispensability like the
352 TK/TKL LECA kinases (Fig. 6). Because the Unaffiliated LECA kinases do not form a monophyletic
353 group in the eukaryotic kinome tree, they should nevertheless not be treated as a set of kinases
354 with similar evolutionary and functional properties. For example, several Unaffiliated LECA kinases
355 including AAK1 and TTK duplicated infrequently but have dispensability levels comparable to LECA
356 kinases from the CMGC superfamily.

357

358 **Discussion**

359 We demonstrate for the first time that by using only slowly evolving kinases, it is possible to
360 generate a eukaryotic kinome tree. The eukaryotic kinome tree was largely automatically
361 annotated into well-supported LECA kinase clades for estimating the ePK repertoire of the LECA.
362 Subsequently, HMM profiles of LECA kinase clades were used to assign the majority of ePKs from
363 94 eukaryotic genomes to the LECA kinase clade from which they most likely originated.

364

365 The eukaryotic kinome tree reveals the phylogenetic relationships between ePKs that were
366 already present in the LECA. The tree can also be used as a platform to better understand ePK
367 evolution on a more functional level. For example, mapping ePK functions on the eukaryotic
368 kinome tree can be used to estimate the relative age of cellular processes that played a role in
369 eukaryogenesis. This requires sufficient internal support to establish the duplication order of LECA
370 kinase clades. However, we observed a puzzling contrast in support between pre-LECA kinase
371 clades (low support) and LECA kinase clades (high support). Interestingly, the limited number of
372 pre-LECA kinase clades that *are* well-supported often include more than two LECA kinase clades.

373

374 Pre-LECA support might improve upon further reducing the BBH set, adding non-kinase domains
375 and advancing phylogenetic methods⁴¹. It is also possible that the exact order of many pre-LECA
376 duplications will remain unsolvable. The existence of well-supported pre-LECA kinase clades that
377 contain multiple LECA kinase clades suggests that these pre-LECA kinase clades may have
378 undergone several rounds of rapid duplication during eukaryogenesis. Rapid duplication
379 complicates reconstructing the duplication order within a pre-LECA kinase clade. An alternative
380 explanation for weak internal support in pre-LECA kinase clades is a syncytial LECA⁴². In that case,
381 presumed pre-LECA duplications are a form of allopolyploidy.

382

383 Based on the eukaryotic kinome tree, we estimated that the number of LECA ePKs is much larger
384 than previously thought. At the same time, our estimate is conservative as only consistent LECA
385 kinase clades with sufficient support values are included. Additional LECA kinase clades might be
386 found upon the sampling of novel eukaryotic species. Especially non-photosynthetic free-living
387 protists are under-sampled in currently sequenced eukaryotes⁴³. The LECA ePK estimate might
388 further increase upon iteratively updating the PFAM HMM profiles Pkinase and Pkinase_Tyr.
389 Searches with the current HMM profiles perhaps resulted in some false negatives⁴⁴ that by
390 inclusion in the BBH sets could result in new LECA kinase clades. Finally, the number of LECA
391 kinase clades could be expanded by manual inspection of potential LECA kinase clades that are not
392 yet included in the current LECA ePK estimate. These potential LECA kinase clades illustrate that
393 our automatic approach is useful at prioritising regions of the eukaryotic kinome tree where
394 manual research is most needed.

395

396 By classifying LECA ePKs in ePK superfamilies, we revealed variation in duplicability and
397 dispensability between and within ePK superfamilies. Orthologous genes that have a wide phyletic
398 distribution are often essential genes^{45,46}. Therefore LECA kinases that display low dispensability
399 may perform functions that have remained essential throughout eukaryotic evolution. LECA
400 kinases that hardly duplicated may also have largely retained their original function from LECA till
401 present-day eukaryotes. Although gene retention or loss after duplication is not necessarily
402 adaptive⁴⁷, highly duplicated genes have been connected to phenotypical changes²¹. Highly
403 duplicated ePKs may thus have contributed to an increased regulatory potential on the
404 evolutionary trajectory from early to present-day eukaryotes. For example, the enormous
405 expansion of kinases from the TK/TKL superfamily in human and other metazoans was indeed

406 essential for the development of multicellularity⁴⁸.

407

408 The CMGC superfamily is unique, as it combines being the largest LECA kinase superfamily with
409 low duplicability post-LECA and low dispensability in present-day species. The current low
410 duplicability and low dispensability are likely due to essential functions that the CMGC kinases
411 performed in the LECA and still perform in present-day species. The large number of duplications
412 in the CMGC superfamily during eukaryogenesis, in contrast, may have allowed adaptive evolution
413 towards the complex eukaryotic cell.

414

415 The eukaryotic kinome tree reveals various new insights in the evolution of kinases. However, the
416 results that we present in this paper also demonstrate that the Scrollsaw method is a valuable
417 approach to generate a well-supported phylogenetic tree starting from a large set of proteins. Our
418 extension of the Scrollsaw method with automatic LECA clade annotation helps to analyse such a
419 tree in a quick and reproducible way. The LECA clade annotation pipeline that we built can, in
420 principle, be applied to all eukaryotic protein domains with sufficient length to generate
421 phylogenetic trees.

422

423 The data that we generated are furthermore a rich resource for a more functional approach to
424 kinases. The eukaryotic kinome tree can serve to select closest neighbours for information
425 transfer, and the overview of LECA kinase retention is useful to select species that are most fit to
426 study the LECA kinase repertoire. The HMM profiles that we provide form a kinome annotation
427 resource on LECA level for newly sequenced eukaryotic species. The first eukaryotic kinome tree is
428 thus relevant both from an evolutionary, methodical and functional perspective.

429

430 **Methods**

431 **Data collection**

432 ***Eukaryotic proteome dataset***

433 In order to collect ePK domains, 94 proteomes were carefully selected from available sequenced
434 eukaryotic genomes. This proteome dataset was compiled as described earlier⁴⁹ but with a slightly
435 different, more diverse set of species. Four fungal species from the earlier dataset were removed
436 while eight new species were added. The proteome dataset contains 1,538,389 proteins in total
437 and is available (see **Data availability**). Details of the selected proteomes can be found in
438 Supplementary Table 3. To each protein in the dataset, a unique protein identifier was assigned
439 that is composed of four letters and six numbers. The four letters combine the first letter of the
440 genus name with the first three letters of the species name. A bifurcating species tree of the
441 species in the eukaryotic proteome dataset was assembled manually. This species tree was rooted
442 between Amorphea and Bikonta¹⁹ and used as input for the LECA clade annotation pipeline.

443

444 ***Kinase domain dataset***

445 From the eukaryotic proteome dataset, kinase domains were selected using PFAM models.
446 All HMM profiles of PFAM-A version 31.0 were downloaded from the PFAM database⁵⁰. These
447 PFAM-A models were used in an HMMSCAN search (HMMER, <http://hmmer.org>, version 3.0)
448 against the eukaryotic proteome dataset (bit score threshold: PFAM Trusted Cutoff). If a particular
449 PFAM-A model hit the same protein sequence multiple times, domain bit scores of non-
450 overlapping hits were summed (a maximum overlap of 30 amino acids was allowed). For each
451 sequence, the best hitting non-overlapping PFAM models were determined based on these
452 modified bit scores. Sequences that were best hit by PFAM models Pkinase and Pkinase_Tyr were
453 collected, and kinase domains were excised based on envelope coordinates. In total, 28,249

453 Pkinase and 8,226 Pkinase_Tyr kinase domains were collected. The resulting fasta file with 36,475
454 kinase domains (Supplementary Data 1) was used as input for the LECA clade annotation pipeline.

455

456 **LECA clade annotation pipeline**

457 The LECA clade annotation pipeline (summarized in Fig. 1) is a Snakemake workflow that consists
458 of a collection of rules in a Snakefile⁵¹. The rules in the Snakefile describe how to create output
459 files from input files. The rules of the LECA clade annotation pipeline Snakefile and their
460 interrelationships are illustrated in a Directed Acyclic Graph (DAG) in Supplementary Fig. 3. To run
461 the LECA clade annotation pipeline, in addition to the Snakefile several data files, scripts and
462 programmes are required. Snakefile, data files and scripts are available (see **Data availability**)
463 while the programmes that were used are listed in the Snakefile. The LECA clade annotation
464 pipeline itself was executed with Snakemake (version 3.11.2). Below, the different steps of the
465 LECA clade annotation pipeline are described, and corresponding Snakefile rules are given in
466 italics.

467

468 ***BBH selection***

469 In order to enable generating a eukaryotic kinome tree, the LECA clade annotation pipeline starts
470 with reducing the kinase domain dataset. The kinase domain dataset was reduced by selecting Bi-
471 directional Best Hits (BBHs) between eukaryotic supergroups. For selecting BBHs, the fasta file
472 with all 36,475 kinase domains (Supplementary Data 1) was divided into separate per species fasta
473 files. These per species kinase fasta files were used in an all species versus all species BLASTp⁵²
474 (version 2.3.0) run (rule *run_species_vs_species_blast*). Based on combining all these BLAST
475 searches, BBHs between eukaryotic supergroups were selected (rules *select_bbh_ids* and
476 *collect_bbh_sequences*). Two different sets of BBHs were compiled. These two different datasets

477 served three purposes: (1) having two different datasets enabled to check results for consistency,
478 (2) a phylogenetic tree with a smaller number of BBHs was easier to annotate with LECA clades
479 while (3) a phylogenetic tree with a larger number of BBHs allowed generating LECA clade HMM
480 profiles with more sequence diversity. The first dataset, the two-supergroups-BBHs dataset,
481 consists of 596 sequences that are BBHs between the two supra-supergroups Amorphea and
482 Bikonta (Supplementary Data 2). The second dataset, the five-supergroups-BBHs dataset, consists
483 of 1,738 sequences that are BBHs between five supergroups that are subsets of either Amorphea
484 ((1) Opisthokonta + Apusozoa and (2) Amoebozoa) or Bikonta ((3) Archaeplastida + Cryptista, (4)
485 SAR (Stramenopiles, Alevolata and Rhizaria) + Haptophyta and (5) Excavata) (Supplementary Data
486 6). The two-supergroups-BBHs dataset is a subset of the five-supergroups-BBHs dataset.

487

488 ***Phylogenetic tree generation***

489 Both the two-supergroups-BBHs dataset and the five-supergroups-BBHs dataset were used to
490 generate a phylogenetic tree. First, both datasets were aligned using mafft-einsi⁵³ (version 7.127)
491 (rule *run_mafft*). Positions in the alignments that did not have a gap score of at least 0.25 were
492 removed with trimAl⁵⁴ (version 1.2rev59) (rule *run_trim_al*). The resulting two-supergroups-BBHs
493 alignment is 263 positions long while the five-supergroups-BBHs alignment contains 261 positions.
494 Alignments were converted to Phylip format (rule *converse_alignment_to_phylip*) and made
495 suitable for RAxML input (rule *prepare_headers_for_raxml*) by changing some characters in the
496 sequence headers (Supplementary Data 3 and 7). Secondly, with RAxML⁵⁵ (version 8.1.1) two
497 maximum likelihood phylogenetic trees were generated using 100 rapid bootstraps, the GAMMA
498 model of rate heterogeneity and an automatically determined best protein substitution model
499 (rule *run_raxml*). For both trees, the best protein substitution model was LG. Annotated versions
500 of the Newick trees and accessory files (see ***Newick file annotation***) are available in

501 Supplementary Data 4, 5, 8 and 9.

502

503 ***LECA clade annotation with Notung***

504 The kinase domain trees were analysed to determine clades (a.k.a. Orthologous Groups (OGs))
505 that were putatively one kinase in the LECA. As a first step to annotate the trees with these LECA
506 clades or LECA OGs, the trees were rearranged with the gene tree-species tree reconciliation
507 software package Notung⁵⁶ (version 2.8.1.6-beta). Notung annotates duplication and speciation
508 nodes in rearranged gene trees. As gene trees are imperfect, Notung was run with two different
509 rearrangement cut-offs. This allowed flexibility in identifying LECA clades.

510

511 Leaves of both the two- and five-supergroups-BBHs trees were prepared for use within Notung by
512 adding supergroup postfixes and species prefixes to leaf names (rules
513 *add_supergroups_to_leaf_names* and *add_species_prefixes_for_notung*). The trees were also
514 midpoint rooted with ETE⁵⁷ (version 3.0.0b29) before running Notung because the default implicit
515 rooting from RAXML could hamper LECA clade annotation at the outer edge of the trees.
516 The two- and five-supergroups-BBHs trees were then rearranged with Notung according to the
517 bifurcating species tree (rule *run_notung*) in the following manner: both trees were rearranged
518 twice, once with bootstrap values below 50 allowed to be rearranged and once with bootstrap
519 values below 70 allowed to be rearranged. The rearranged trees were stripped of species prefixes
520 because the prefixes were redundant after running Notung (rule *remove_species_prefixes*).

521

522 In the four rearranged Notung trees, LECA clades were determined using a custom script that
523 started with pre-LECA duplication nodes (rule *determine_notung_ogs*). Pre-LECA duplication nodes
524 represent gene duplications that preceded all species and thus occurred before the LECA

525 originated. They were parsed from a list of duplication nodes that Notung offers as output. In the
526 rearranged phylogenetic trees, nodes that are children of pre-LECA duplication nodes but are not
527 pre-LECA duplication nodes themselves were assessed as potential LECA speciation nodes.
528 Potential LECA speciation nodes with at least one Amorphea and one Bikonta sequence among
529 their children were defined as definitive LECA speciation nodes. All sequences descending from a
530 definitive LECA speciation node were then defined as forming a single LECA clade. The reliability of
531 a LECA clade that was annotated in a rearranged Notung tree was determined by evaluating the
532 corresponding original RAxML tree. A LECA clade was regarded reliable if all sequences belonging
533 to it also formed a single clade in the original RAxML tree. LECA clades that were not monophyletic
534 in the original RAxML trees were labelled dubious and removed (rule
535 *determine_dubious_notung_ogs*).

536

537 Because two different bootstrap cut-offs (50 and 70) were used to generate rearranged Notung
538 trees, two different sets of LECA clades were available per original RAxML tree. For each RAxML
539 tree, the two sets of LECA clades were combined into one set that attempted to annotate as many
540 tree leaves as possible (rule *combine_notung_ogs*). The rationale behind maximizing the number
541 of annotated leaves is that each present-day kinase likely originates from a LECA kinase clade.
542 Preferably, LECA clades based on the most stringent rearrangement cutoff 70 (70-clades) were
543 used for leaf annotation because 70-clades are better supported than clades based on
544 rearrangement cutoff 50 (50-clades). However, 70-clades that were labelled as dubious or had
545 bootstrap support below 50 were not used in the two combined sets of LECA clades. If possible,
546 they were replaced by one or more 50-clades. Only non-dubious 50-clades with minimal bootstrap
547 support of 50 could serve as a replacement. For the two-supergroups-BBHs tree, the combined set
548 of LECA clades based on Notung consists of 117 LECA clades that annotate 535 of the 596 leaves.

549 For the five-supergroups-BBHs tree, the combined set consists of 113 LECA clades that annotate
550 1,283 of the 1,738 leaves.

551

552 ***LECA clade annotation with HMMER***

553 In a second tree annotation step, LECA clades annotated with Notung were expanded with not yet
554 annotated tree leaves using HMMER. After annotating putative LECA clades with Notung, several
555 sequences in the trees were not annotated, even though they resided close by annotated Notung
556 clades. These leaves were prevented from being part of a Notung LECA clade by bootstrap values
557 or inconsistencies between gene tree and species tree. To still annotate these leaves with existing
558 Notung LECA clades, the two combined sets of Notung LECA clades were projected on the two BBH
559 sets in two consecutive rounds of HMMER searches.

560

561 Notung LECA clade HMM profiles for the first HMMER search against BBHs were generated as
562 follows. For both the two- and five-supergroups-BBHs, BBH sequences that were annotated as part
563 of a Notung LECA clade were collected for each Notung LECA clade in a fasta file (rule
564 *distribute_ogs_1*). Each Notung LECA clade fasta file was aligned with mafft-einsi, and the
565 alignment was used to generate a HMMER3 profile. All two-supergroups-BBHs Notung LECA clade
566 HMMER3 profiles were used for a HMMER search against the two-supergroups-BBH sequences
567 (rule *run_hmmer_search_bbhs_1*). The five-supergroups-BBHs Notung LECA clade HMMER3
568 profiles were used for a similar search against the five-supergroups-BBH sequences. A list of top
569 two best hitting Notung LECA clade HMM profiles was generated for both the two- and five-
570 supergroups-BBHs (rule *determine_hmmer_ogs_bbhs_1*). Only BBHs with a bit score difference of
571 minimal 10 between top two hits and at least one bit score of minimal 30 were listed. Leaves of
572 the two- and five-supergroups-BBH trees that were not yet annotated with Notung were then

573 annotated with the Notung LECA clade that was their best HMMER hit in the top two list (rule
574 *add_hmmer_ogs_bbhs_1*). Leaf annotation with the best Notung LECA clade HMMER hit was only
575 completed provided the Notung LECA clade leaves and the leaf best hit by the associated Notung
576 LECA clade HMMER profile were monophyletic in the rooted RAXML tree. In the first round of
577 HMMER annotation, 42 leaves of the two-supergroups-BBHs tree and 167 leaves of the five-
578 supergroups-BBHs tree were annotated with HMMER in addition to the Notung annotation.
579
580 The expansion of Notung LECA clades with leaves annotated with HMMER allowed generating
581 more sensitive HMMER profiles. These more sensitive HMMER profiles for the second HMMER
582 search against BBHs were generated as follows. BBH sequences that were annotated with Notung
583 or HMMER as forming a single LECA clade were combined, aligned with mafft-einsi and
584 subsequently a Notung-HMMER LECA clade HMMER3 profile was generated (rule
585 *distribute_ogs_2*). All two-supergroups-BBHs Notung-HMMER LECA clade HMMER3 profiles were
586 combined into one set, and the same was done for all five-supergroups-BBHs Notung-HMMER
587 LECA clade HMMER3 profiles. These two sets of HMMER profiles were each used for a HMMER
588 search against both the two- and five-supergroups-BBH sequences (rule
589 *run_hmmer_search_bbhs_2*). The HMMER searches with HMMER profiles corresponding to their
590 source tree were used for further leaf annotation (e.g. two-supergroups-BBHs Notung-HMMER
591 LECA clade profiles against two-supergroups-BBHs). The HMMER searches with HMMER profiles
592 derived from the other tree were later in the pipeline used for mapping leaf annotation between
593 the two- and five-supergroups-BBH trees (e.g. two-supergroups-BBHs Notung-HMMER LECA clade
594 profiles against five-supergroups-BBHs). For each of the four HMMER searches a list of each BBHs
595 top two best hitting Notung-HMMER LECA clade HMMER profiles was generated (rule
596 *determine_hmmer_ogs_bbhs_2*). This was done under the same conditions as described for the

597 first round of HMMER annotation. Leaves of the two- and five-supergroups-BBH trees that were
598 not yet annotated with Notung or the first round of HMMER annotation were then annotated with
599 the best hitting Notung-HMMER LECA clade from their respective Notung-HMMER LECA clade set
600 (rule *add_hmmer_ogs_bbhs_2*). Leaf annotation was again only completed provided the Notung-
601 HMMER LECA clade leaves and the leaf best hit by the associated Notung-HMMER LECA clade
602 HMMER profile were monophyletic in the rooted RAXML tree. In this second round of HMMER
603 annotation, three leaves of the two-supergroups-BBHs tree and 49 leaves of the five-supergroups-
604 BBHs tree were annotated on top of the existing annotation. In total, 580 of the 596 leaves of the
605 two-supergroups-BBHs tree and 1,499 of the 1,738 leaves of the five-supergroups-BBHs tree were
606 annotated. Notung-HMMER LECA clade sequences and sequences newly annotated with the same
607 LECA clade in the second round of HMMER searches were combined in per LECA clade fasta files
608 (rule *distribute_ogs_3*).

609

610 ***Combination LECA clades two- and five-supergroups-BBHs trees***

611 In a third step to annotate the trees with LECA clades, Notung-HMMER LECA clades of the two-
612 and five-supergroups-BBHs trees were combined into one overarching set. To do this, first for each
613 tree a list with leaves was generated that per leaf provides the two Notung-HMMER LECA clades
614 to which the leaf is annotated in respectively the two- and five-supergroups-BBHs trees (rule
615 *map_ogs*). This list was subsequently used to extract which Notung-HMMER LECA clade in the
616 two-supergroups-BBHs tree corresponds to which Notung-HMMER LECA clade in the five-
617 supergroups-BBHs tree and vice versa. If leaves of a Notung-HMMER LECA clade in one of the two
618 trees were distributed over multiple Notung-HMMER LECA clades in the other tree, these multiple
619 LECA clades were merged into a single LECA clade (rule *merge_ogs*) (Supplementary Table 4). After
620 merging, the number of Notung-HMMER LECA clades annotated in the two-supergroups-BBHs tree

621 decreased from 117 to 110.

622

623 Based on the mapping of Notung-HMMER LECA clades between the two- and five-supergroups-
624 BBHs trees and the merged LECA clades, a new set of combined LECA clades was generated (rule
625 *combine_ogs*). The 110 Notung-HMMER LECA clades that were annotated in the two-supergroups-
626 BBHs tree and the 113 Notung-HMMER LECA clades that were annotated in the five-supergroups-
627 BBHs tree formed together 118 unique Notung-HMMER LECA clades. Eight Notung-HMMER LECA
628 clades were absent in the two-supergroups-BBHs tree, and five Notung-HMMER LECA clades were
629 not automatically annotated in the five-supergroups-BBHs tree (Supplementary Table 5). Per
630 combined LECA clade, sequences of the Notung-HMMER LECA clade(s) that form the combined
631 LECA clade were collected, aligned with mafft-einsi and used to generate a HMMER3 profile (rule
632 *combine_og_profiles*). A modified version of these HMMER3 profiles (see **Manual Annotation**) is
633 available in Supplementary Data 10.

634

635 **Domain assignment**

636 The combined set of 118 unique Notung-HMMER LECA clades was used to assign the initial 36,475
637 kinase domains to a LECA clade. In order to do this, the combined LECA clade HMMER3 profiles
638 were run against the complete kinase domain dataset (rule *run_hmmer_search_all_4*). The results
639 of this HMMER run were used to divide kinase domains over three lists (rule
640 *determine_hmmer_ogs_all_4*): (1) assigned kinase domains with a bit score difference of minimal
641 10 between top two LECA clade hits and at least one LECA clade hit with a bit score of minimal 30,
642 (2) difficult-to-assign kinase domains with a bit score difference below 10 between top two LECA
643 clade hits and (3) unassigned kinase domains with only LECA clade hits with a bit score below 30.
644 Modified versions of these lists (see **Manual Annotation**) are available as Supplementary Data 11-

645 13. The assigned kinase domains from the first list were used to generate a matrix that for each
646 species in the eukaryotic proteome dataset and for each combined LECA clade provides how many
647 sequences were hit. An adjusted version of this matrix (Supplementary Data 14) forms the basis
648 for Fig. 3 to 6. Furthermore, the assigned kinase domains were classified per eukaryotic
649 supergroup to determine the number of eukaryotic supergroups hit by each LECA clade. A
650 supergroup was only counted if kinases from minimal two species of that supergroup were hit by a
651 particular combined LECA clade. Per species, the percentage of assigned kinase domains was also
652 determined (rule *determine_assignment_percentages*) (Supplementary Table 2).

653

654 **LECA clade categorization**

655 The reliability of the set of 118 combined LECA clades was evaluated using information directly
656 and indirectly available in the LECA clade annotation pipeline. The combined LECA clades were
657 therefore classified in categories with respect to the amount of support they have from the two
658 different trees and domain assignment. Per combined LECA clade, the following information was
659 listed (rule *add_hmmer_supergroups*): (1) presence/absence in both RAxML trees, (2) number of
660 eukaryotic supergroups among assigned kinases using the counting mode of the supergroup
661 matrix described under **Domain assignment**, (3) bootstrap support in both RAxML trees, and (4)
662 correspondence to multiple Notung-HMMER LECA clades in one of the RAxML trees. Based on this
663 information, LECA clades were labelled with four categories (rule *add_categories*): (1) combined
664 LECA clades that are annotated only in one of the two trees (indicated with *), (2) combined LECA
665 clades to which domains from less than two eukaryotic supergroups were assigned with HMMER
666 (indicated with **), (3) combined LECA clades that in neither of the RAxML trees have bootstrap
667 support of minimal 70 (indicated with ***) and (4) combined LECA clades that in one of the trees
668 are split in multiple Notung-HMMER LECA clades (indicated with %). A fifth category was added

669 later upon manual annotation (see **Manual annotation**). Not all LECA clades are labelled with any
670 of the categories while LECA clades can also be labelled with multiple categories at once.

671

672 **LECA clade naming**

673 The 118 combined LECA clades were named in order to distinguish them better and easily link
674 them with functional information (rule *make_og_name_table_and_list*). They were named by
675 their best hit in the well-studied species human, baker's yeast or *A. thaliana*. Preferably, human
676 kinases were used for naming, but if no human kinase was assigned to a LECA clade, baker's yeast
677 or *A. thaliana* names were used. If hits from all three species were absent, a LECA clade was
678 indicated with its combined LECA clade number. Furthermore, all kinases from human, baker's
679 yeast and *A. thaliana* that were assigned to a LECA clade were listed in a table (Supplementary
680 Table 1). Per LECA clade, hits from these species were displayed in descending bit score order.
681 LECA clades in the table were extended with categories (rule *add_og_categories_to_table*). LECA
682 clade names were also added to earlier generated files (rules *add_og_names_to_list* and
683 *add_og_names_to_matrix*).

684 *The*

685 **Newick file annotation**

686 To browse through the eukaryotic kinome tree and LECA clade annotation easily at once, leaf
687 names of the Newick files of the two- and five-supergroups-BBHs trees were extended with LECA
688 clade annotation. Leaf names had already been extended with supergroup names before (rule
689 *add_supergroups_to_leaf_names*). If leaves were annotated with Notung of HMMER, they were
690 extended with a Notung LECA clade (abbreviated as nOG) (rule *add_notung_ogs_to_leaf_names*)
691 or HMMER LECA clade (abbreviated as hOG) (rules *add_hmmer_ogs_to_leaf_names_bbhs_1* and
692 *add_hmmer_ogs_to_leaf_names_bbhs_2*). Subsequently, leaves that participated in a combined

693 LECA clade were extended with this combined LECA clade (abbreviated as cOG) including both its
694 number, categories and name (rule *add_combined_ogs_to_leaf_names*). Leaves from combined
695 LECA clades that include manually annotated leaves (see **Manual annotation**) were denoted as
696 mOG instead of cOG. Finally, to all leaves the top two best hitting combined LECA clades were
697 added together with their bit scores (rule *add_combined_og_hmmer_hits_to_leaf_names*).
698 For visualising the trees with iTOL⁵⁸, nodes that are the common ancestor of leaves that form a
699 LECA clade were named with this LECA clade (rule *add_ogs_to_nodes*). Furthermore, to
700 automatically collapse the LECA clades in iTOL, a 'collapse file' was generated. Annotated Newick
701 files of the two- and five-supergroups-BBHs trees (Supplementary Data 4 and 8) and their collapse
702 files (Supplementary Data 5 and 9) form the basis for Fig. 2 and Supplementary Fig. 1.

703

704 **Manual annotation**

705 When inspecting the annotated Newick trees, the automatic LECA clade annotation displayed
706 room for manual improvement. Manual annotation was performed in the following cases: (1) to
707 split merged LECA clades (indicated with %) if there were good reasons to believe that they are
708 indeed multiple LECA clades, (2) to merge LECA clades if there were good reasons to believe that
709 they are indeed one LECA clade or (3) to annotate not yet annotated leaves. Manual annotation
710 was done by partially re-executing the LECA clade annotation pipeline (starting with rule
711 *run_hmmer_search_all_4*) after copying and manually changing files including the HMMER3
712 profiles of combined LECA clades (rule *copy_lists_notung_hmmer_hmmer_ogs* and the description
713 of manual runs in the Snakefile). Manual annotation occurred in two rounds, with annotated
714 Newick trees of the first manual run serving to determine the next step in the second manual run.
715 The LECA clades that include manually annotated leaves and the reasoning that resulted in their
716 manual annotation are described in Supplementary Table 6. In total, 16 LECA clades were fully or

717 partially based on manually annotated leaves.

718

719 The total number of combined LECA clades did not change after manual annotation and remained
720 118. But the number of annotated combined LECA clades in the two-supergroups-BBHs tree
721 increased from 110 to 113 and the number of annotated combined LECA clades in the five-
722 supergroups-BBHs tree decreased from 113 to 111 (Supplementary Table 6).

723

724 Manually annotated LECA clades form a fifth category (see *LECA clade categorization*) that is
725 indicated with #. In the Newick trees, their leaves are indicated with mOG instead of cOG. In the
726 two-supergroups-BBHs tree, 17 of the 596 leaves were manually annotated resulting in the final
727 annotation of 593 leaves. In the five-supergroups-BBHs tree, 264 of the 1,738 leaves were
728 manually annotated resulting in the final annotation of 1,585 leaves.

729

730 **Figure generation**

731 Fig. 1 was produced with Lucidchart, <https://www.lucidchart.com>. Fig. 2 was produced with iTOL⁵⁸
732 (version 4.3) based on Supplementary Data 4 and 5. Fig. 3 to 6 were produced with R⁵⁹ (version
733 3.4.4) based on Supplementary Data 14 and 16, using R packages APE⁶⁰ and gplots⁶¹. All figures
734 were adjusted with Inkscape, <http://inkscape.org>.

735

736 **Data availability**

737 The eukaryotic proteome dataset is available at
738 https://bioinformatics.bio.uu.nl/snel/support/eukaryotic_proteome_dataset. The entire LECA
739 clade annotation pipeline, including all input data and output files, is available at
740 https://bioinformatics.bio.uu.nl/snel/support/LECA_clade_annotation_pipeline. A selection of

741 output files is also available as Supplementary Data.

742

743 **Code availability**

744 The computational code of the LECA clade annotation pipeline is together with the data available

745 at https://bioinformatics.bio.uu.nl/snel/support/LECA_clade_annotation_pipeline.

746

747 **References**

- 748 1. Lahiry, P., Torkamani, A., Schork, N. J. & Hegele, R. A. Kinase mutations in human disease:
749 interpreting genotype-phenotype relationships. *Nat. Rev. Genet.* **11**, 60–74 (2010).
- 750 2. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature reviews. Drug discovery* **1**,
751 727–730 (2002).
- 752 3. Merritt, C., Silva, L. E., Tanner, A. L., Stuart, K. & Pollastri, M. P. Kinases as druggable targets
753 in trypanosomatid protozoan parasites. *Chem. Rev.* **114**, 11280–11304 (2014).
- 754 4. Spitzmuller, A. & Mestres, J. Prediction of the *P. falciparum* target space relevant to malaria
755 drug discovery. *PLoS Comput. Biol.* **9**, e1003257 (2013).
- 756 5. Lee, K.-T. *et al.* Systematic functional analysis of kinases in the fungal pathogen
757 *Cryptococcus neoformans*. *Nat. Commun.* **7**, 12766 (2016).
- 758 6. Kulik, A., Wawer, I., Krzywinska, E., Bucholc, M. & Dobrowolska, G. SnRK2 protein kinases--
759 key regulators of plant response to abiotic stresses. *OMICS* **15**, 859–872 (2011).
- 760 7. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase
761 complement of the human genome. *Science* **298**, 1912–34 (2002).
- 762 8. Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C. & Manning, G. Structural and functional
763 diversity of the microbial kinome. *PLoS Biol.* **5**, e17 (2007).
- 764 9. Hanks, S. K. & Hunter, T. Protein kinases 6. The eukaryotic protein kinase superfamily:

- 765 kinase (catalytic) domain structure and classification. *FASEB J. Off. Publ. Fed. Am. Soc. Exp.*
766 *Biol.* **9**, 576–596 (1995).
- 767 10. Bradham, C. A. *et al.* The sea urchin kinome: a first look. *Dev. Biol.* **300**, 180–193 (2006).
- 768 11. Goldberg, J. M. *et al.* The dictyostelium kinome--analysis of the protein kinases from a
769 simple model organism. *PLoS Genet.* **2**, e38 (2006).
- 770 12. Bemm, F., Schwarz, R., Forster, F. & Schultz, J. A kinome of 2600 in the ciliate *Paramecium*
771 *tetraurelia*. *FEBS Lett.* **583**, 3589–3592 (2009).
- 772 13. Talevich, E., Tobin, A. B., Kannan, N. & Doerig, C. An evolutionary perspective on the kinome
773 of malaria parasites. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 2607–2618 (2012).
- 774 14. Zulawski, M., Schulze, G., Braginets, R., Hartmann, S. & Schulze, W. X. The Arabidopsis
775 Kinome: phylogeny and evolutionary insights into functional diversification. *BMC Genomics*
776 **15**, 548 (2014).
- 777 15. Lynch, M. *et al.* Evolutionary cell biology: two origins, one objective. *Proc. Natl. Acad. Sci. U.*
778 *S. A.* **111**, 16990–16994 (2014).
- 779 16. Manning, G. *et al.* The minimal kinome of *Giardia lamblia* illuminates early kinase evolution
780 and unique parasite biology. *Genome Biol.* **12**, R66 (2011).
- 781 17. Klopstein, S., Massingham, T. & Goldman, N. More on the Best Evolutionary Rate for
782 Phylogenetic Analysis. *Syst. Biol.* **66**, 769–785 (2017).
- 783 18. Elias, M., Brighouse, A., Gabernet-Castello, C., Field, M. C. & Dacks, J. B. Sculpting the
784 endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J. Cell*
785 *Sci.* **125**, 2500–8 (2012).
- 786 19. Burki, F. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring*
787 *Harb. Perspect. Biol.* **6**, a016147 (2014).
- 788 20. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of

- 789 eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).
- 790 21. Lehti-Shiu, M. D. & Shiu, S.-H. Diversity, classification and function of the plant protein
791 kinase superfamily. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 2619–2639 (2012).
- 792 22. Higgins, J. M. Haspin-like proteins: a new family of evolutionarily conserved putative
793 eukaryotic protein kinases. *Protein Sci.* **10**, 1677–1684 (2001).
- 794 23. Lizcano, J. M. *et al.* LKB1 is a master kinase that activates 13 kinases of the AMPK subfamily,
795 including MARK/PAR-1. *EMBO J.* **23**, 833–843 (2004).
- 796 24. Hawley, S. A. *et al.* Calmodulin-dependent protein kinase kinase-beta is an alternative
797 upstream kinase for AMP-activated protein kinase. *Cell Metab.* **2**, 9–19 (2005).
- 798 25. Crozet, P. *et al.* Mechanisms of regulation of SNF1/AMPK/SnRK1 protein kinases. *Front.*
799 *Plant Sci.* **5**, 1–17 (2014).
- 800 26. Anderson, K. A. *et al.* Components of a calmodulin-dependent protein kinase cascade.
801 Molecular cloning, functional characterization and cellular localization of Ca²⁺/calmodulin-
802 dependent protein kinase kinase beta. *J. Biol. Chem.* **273**, 31880–31889 (1998).
- 803 27. Mora, A., Komander, D., van Aalten, D. M. F. & Alessi, D. R. PDK1, the master regulator of
804 AGC kinase signal transduction. *Semin. Cell Dev. Biol.* **15**, 161–170 (2004).
- 805 28. Batistic, O. & Kudla, J. Plant calcineurin B-like proteins and their interacting protein kinases.
806 *Biochim. Biophys. Acta* **1793**, 985–992 (2009).
- 807 29. Chen, Y. & Sanchez, Y. Chk1 in the DNA damage response: conserved roles from yeasts to
808 mammals. *DNA Repair (Amst)*. **3**, 1025–1032 (2004).
- 809 30. Liu, J., Ishitani, M., Halfter, U., Kim, C. S. & Zhu, J. K. The Arabidopsis thaliana SOS2 gene
810 encodes a protein kinase that is required for salt tolerance. *Proc. Natl. Acad. Sci. U. S. A.* **97**,
811 3730–3734 (2000).
- 812 31. Ohta, M., Guo, Y., Halfter, U. & Zhu, J.-K. A novel domain in the protein kinase SOS2

- 813 mediates interaction with the protein phosphatase 2C ABI2. *Proc. Natl. Acad. Sci. U. S. A.*
814 **100**, 11771–6 (2003).
- 815 32. Gong, E.-Y. *et al.* KA1-targeted regulatory domain mutations activate Chk1 in the absence of
816 DNA damage. *Sci. Rep.* **5**, 10856 (2015).
- 817 33. Beckmann, L., Edel, K. H., Batistič, O. & Kudla, J. A calcium sensor – protein kinase signaling
818 module diversified in plants and is retained in all lineages of Bikonta species. *Sci. Rep.* **6**,
819 31645 (2016).
- 820 34. KinBase. Available at: <http://kinase.com/web/current/kinbase>. (Accessed: 3rd May 2019)
- 821 35. Aranda, S., Laguna, A. & de la Luna, S. DYRK family of protein kinases: evolutionary
822 relationships, biochemical properties, and functional roles. *FASEB J. Off. Publ. Fed. Am. Soc.*
823 *Exp. Biol.* **25**, 449–462 (2011).
- 824 36. Glenewinkel, F. *et al.* The adaptor protein DCAF7 mediates the interaction of the adenovirus
825 E1A oncoprotein with the protein kinases DYRK1A and HIPK2. *Sci. Rep.* **6**, 28241 (2016).
- 826 37. Rodriguez-Gil, A. *et al.* HIPK family kinases bind and regulate the function of the CCR4-NOT
827 complex. *Mol. Biol. Cell* **27**, 1969–1980 (2016).
- 828 38. Nakjang, S. *et al.* Reduction and expansion in microsporidian genome evolution: new
829 insights from comparative genomics. *Genome Biol. Evol.* **5**, 2285–2303 (2013).
- 830 39. Lee, J. *et al.* Analysis of the Draft Genome of the Red Seaweed *Gracilariopsis chorda*
831 Provides Insights into Genome Size Evolution in Rhodophyta. *Mol. Biol. Evol.* **35**, 1869–1886
832 (2018).
- 833 40. Deguchi, A. *et al.* LKB1 suppresses p21-activated kinase-1 (PAK1) by phosphorylation of
834 Thr109 in the p21-binding domain. *J. Biol. Chem.* **285**, 18283–18290 (2010).
- 835 41. Lemoine, F. *et al.* Renewing Felsenstein’s phylogenetic bootstrap in the era of big data.
836 *Nature* **556**, 452–456 (2018).

- 837 42. Garg, S. G. & Martin, W. F. Mitochondria, the Cell Cycle, and the Origin of Sex via a Syncytial
838 Eukaryote Common Ancestor. *Genome Biol. Evol.* **8**, 1950–1970 (2016).
- 839 43. del Campo, J. *et al.* The others: our biased perspective of eukaryotic genomes. *Trends Ecol.*
840 *Evol.* **29**, 252–259 (2014).
- 841 44. Menichelli, C., Gascuel, O. & Brehelin, L. Improving pairwise comparison of protein
842 sequences with domain co-occurrence. *PLoS Comput. Biol.* **14**, e1005889 (2018).
- 843 45. Waterhouse, R. M., Zdobnov, E. M. & Kriventseva, E. V. Correlating traits of gene retention,
844 sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi.
845 *Genome Biol. Evol.* **3**, 75–86 (2011).
- 846 46. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene
847 essentiality. *Nat. Rev. Genet.* **19**, 34–49 (2018).
- 848 47. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing
849 between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
- 850 48. Suga, H., Torruella, G., Burger, G., Brown, M. W. & Ruiz-Trillo, I. Earliest Holozoan expansion
851 of phosphotyrosine signaling. *Mol. Biol. Evol.* **31**, 517–528 (2014).
- 852 49. van Hooff, J. J., Tromer, E., van Wijk, L. M., Snel, B. & Kops, G. J. Evolutionary dynamics of
853 the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep.*
854 **18**, 1559–1571 (2017).
- 855 50. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.
856 *Nucleic Acids Res.* **44**, D279–85 (2016).
- 857 51. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine.
858 *Bioinformatics* **28**, 2520–2522 (2012).
- 859 52. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
860 (2009).

- 861 53. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
862 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–80 (2013).
- 863 54. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated
864 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
865 (2009).
- 866 55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
867 phylogenies. *Bioinformatics* **30**, 1312–3 (2014).
- 868 56. Stolzer, M. *et al.* Inferring duplications, losses, transfers and incomplete lineage sorting with
869 nonbinary species trees. *Bioinformatics* **28**, i409–i415 (2012).
- 870 57. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of
871 Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
- 872 58. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new
873 developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
- 874 59. R Core Team (2019). R: A language and environment for statistical computing. R Foundation
875 for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- 876 60. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R
877 language. *Bioinformatics* **20**, 289–290 (2004).
- 878 61. Warnes, G. R., *et al.* gplots: Various R programming tools for plotting data. R package
879 version 3.0.1. <https://cran.r-project.org/web/packages/gplots/gplots.pdf>.

880

881 **Acknowledgements**

882 We thank Sebastiaan Broekema for carrying out a pilot project on the usability of the ScrollSaw
883 method for the eukaryotic kinome, John van Dam for his contribution to compiling the eukaryotic
884 proteome dataset and members of the Snel lab for critical reading and helpful discussion on the

885 manuscript. We thank Ivica Letunic for support with iTOL. This research was financially supported
886 by the Netherlands Organization for Scientific Research (NWO) grant 016.160.638 Vici.

887

888 **Author information**

889 **Affiliations**

890 *Theoretical Biology and Bioinformatics, Department of Biology, Science Faculty, Utrecht University,*
891 *Utrecht, The Netherlands*

892 B. Snel, L.M. van Wijk

893

894 **Contributions**

895 B.S. and L.M.W. designed the research. L.M.W. performed the research and analysed the data. B.S.
896 and L.M.W. wrote the manuscript.

897

898 **Corresponding author**

899 Correspondence to B. Snel.

900

901 **Ethics declarations**

902 **Competing interests**

903 The authors declare no competing interests.

904

905 **Supplementary material**

906 Supplementary Information:

907 Description of Supplementary Tables

908 Description of Supplementary Data

909 Supplementary Results

910 Supplementary References

911 Supplementary Fig. 1-3

912 Supplementary Tables 1-16

913 Supplementary Data 1-16

914

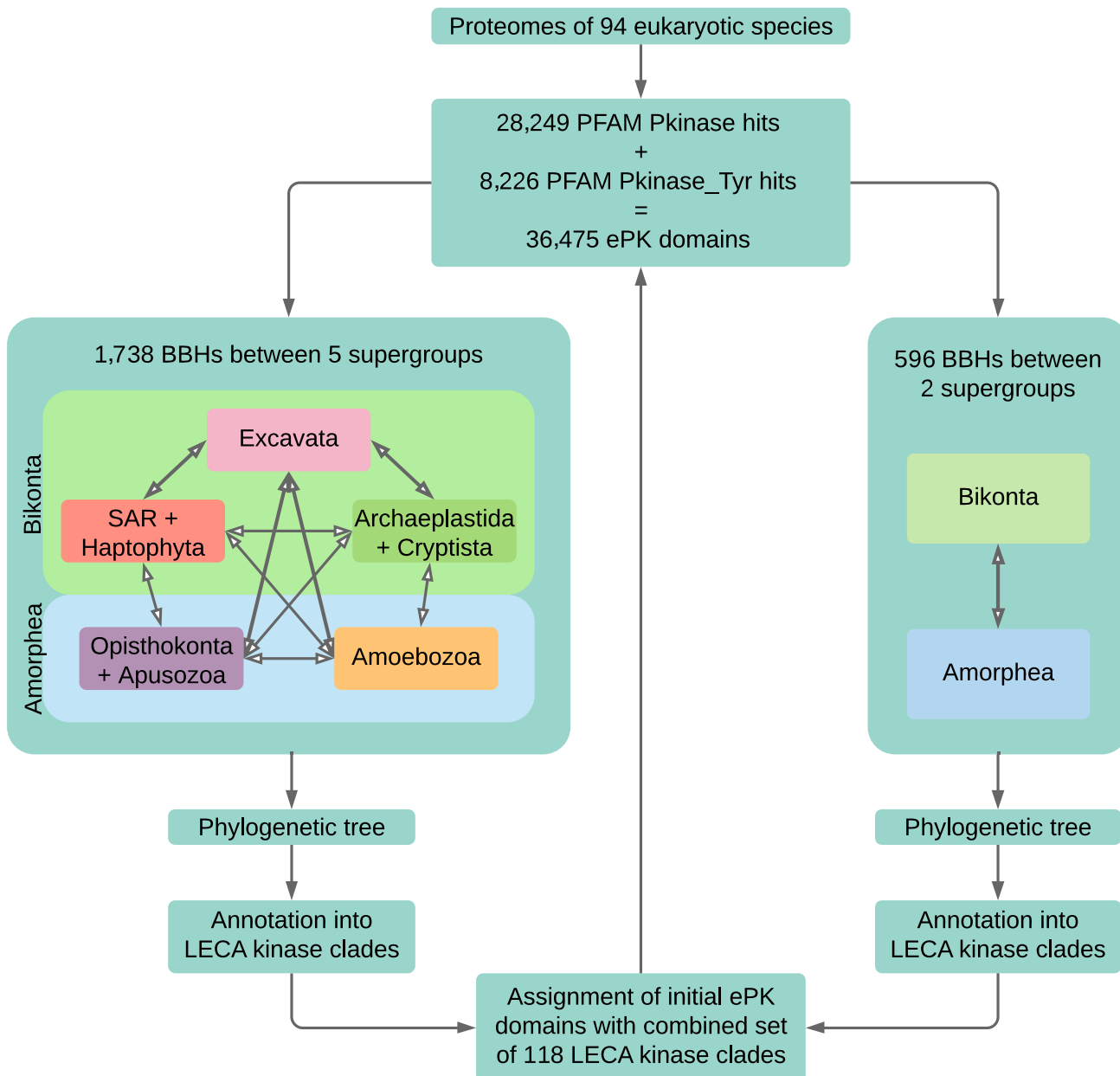


Fig. 1: Overview of important steps in the LECA clade annotation pipeline.

Superfamily:

- UNAF
- CAMK
- AGC
- CMGC
- STE
- CK1
- TK/TKL



Fig. 2: The eukaryotic kinome tree based on two-supergroups-BBHs. LECA kinase clades are colour coded according to ePK superfamily. LECA kinase clade names indicated with SCER or ATHA in-between brackets are not derived from human but from baker's yeast or *A. thaliana* protein names, respectively. Unaffiliated LECA kinase clades are grey. LECA kinase clades that fail one or more criteria for inclusion in the LECA kinase number estimate are indicated with *(absence in one of the trees), ** (limited distribution over species) and *** (bootstrap support below 70 in both trees). LECA kinase clades that initially were split into two LECA kinase clades in one of the trees are indicated with %. LECA kinase clades that include manually annotated leaves are indicated with #. Bootstrap support of minimal 50 out of 100 is shown.

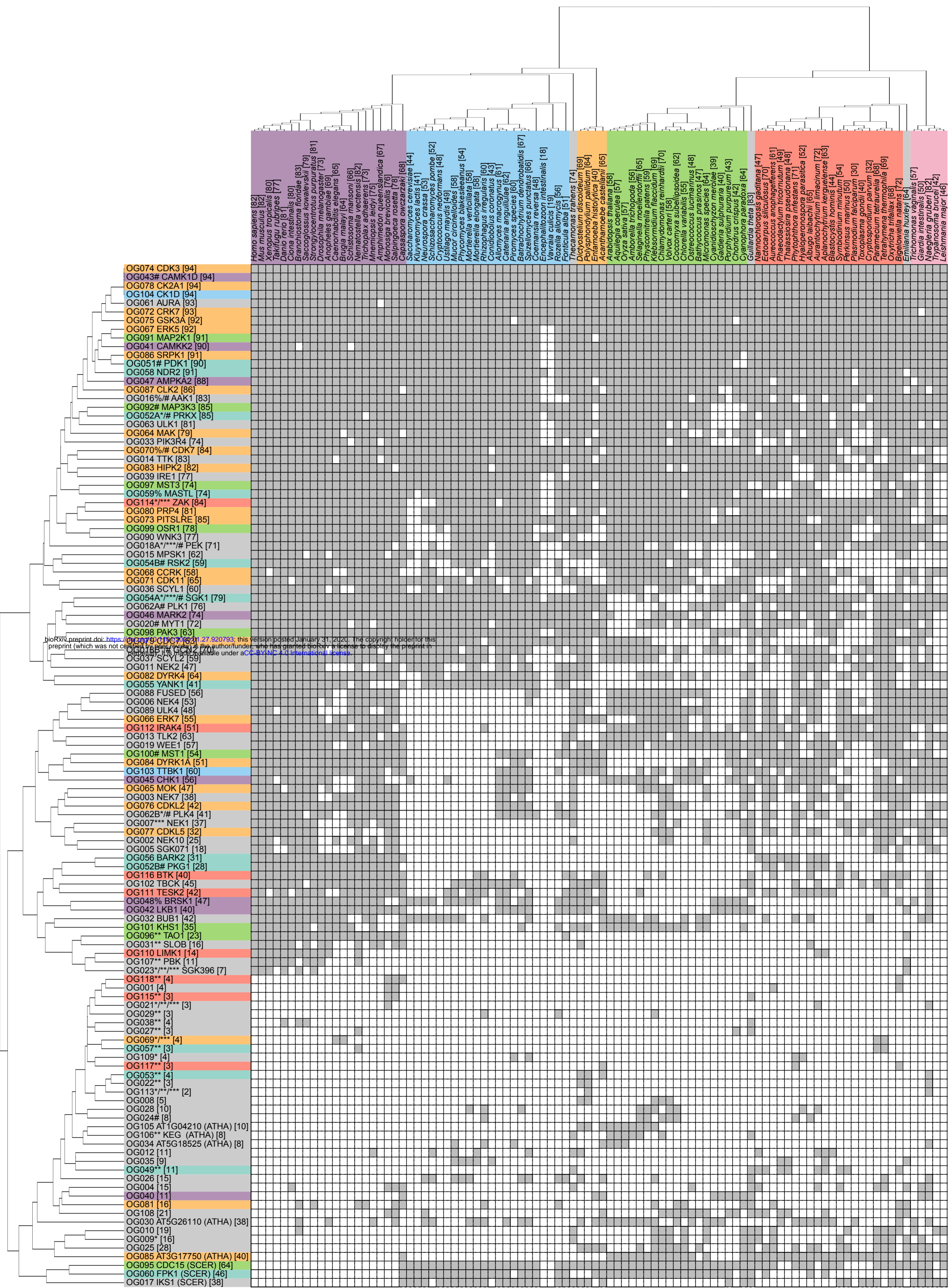


Fig. 3: Clustered presence/absence profile of 28,893 ePK domains in 94 present-day eukaryotes. Assignment of one or more ePKs from a species to a certain LECA kinase clade is indicated in grey while absence is indicated in white. LECA kinase colour code and the meaning of special characters is the same as in Fig. 2. Species are colour coded according to eukaryotic supergroup as in Fig. 4. Behind LECA kinases the total number of species from which at least one ePK was assigned to a particular LECA kinase clade is given. Behind species names the total number of LECA kinase clades to which ePKs from a particular species were assigned is given.

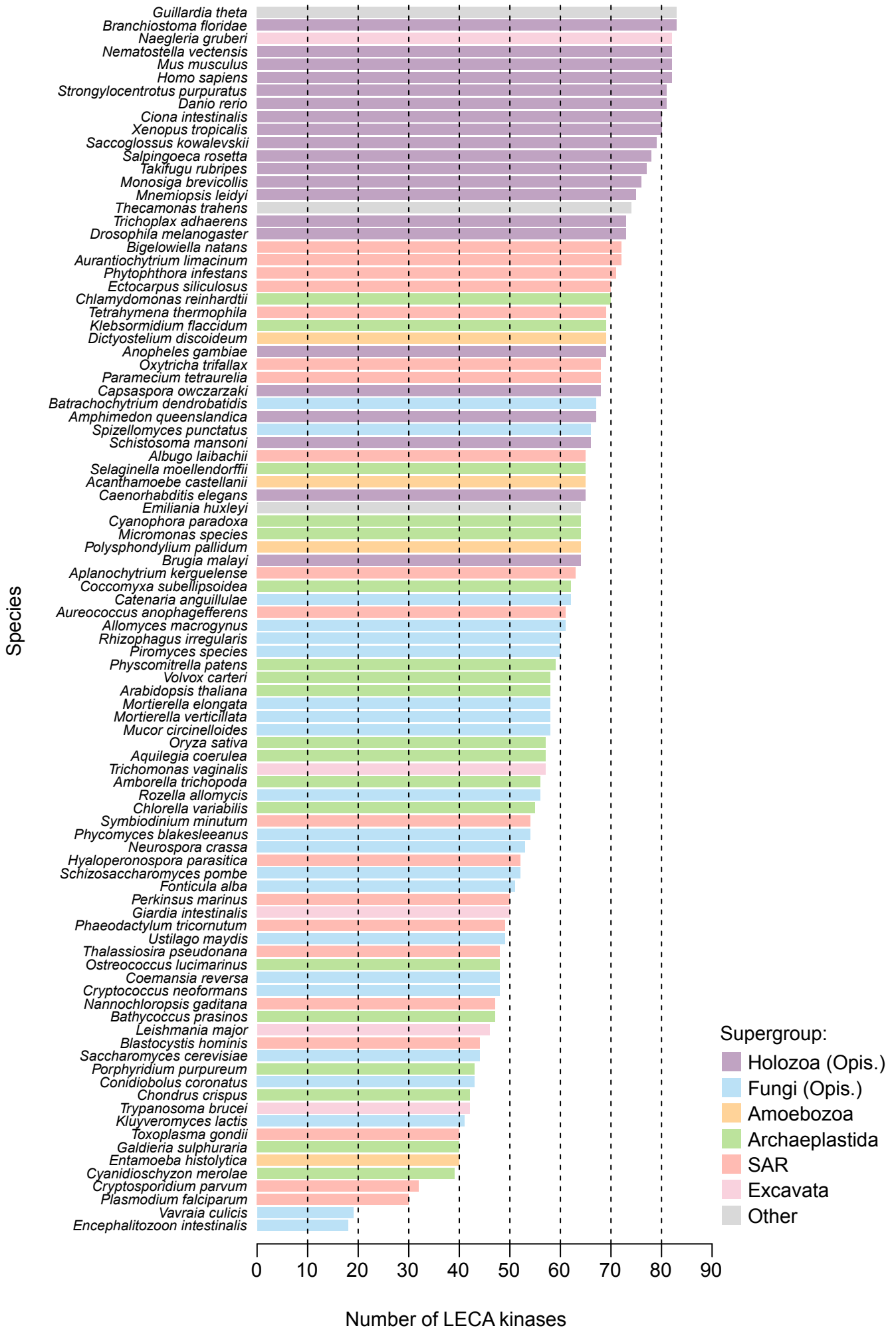


Fig. 4: LECA kinase retention in 94 present-day eukaryotes. Species are colour coded according to eukaryotic supergroup.

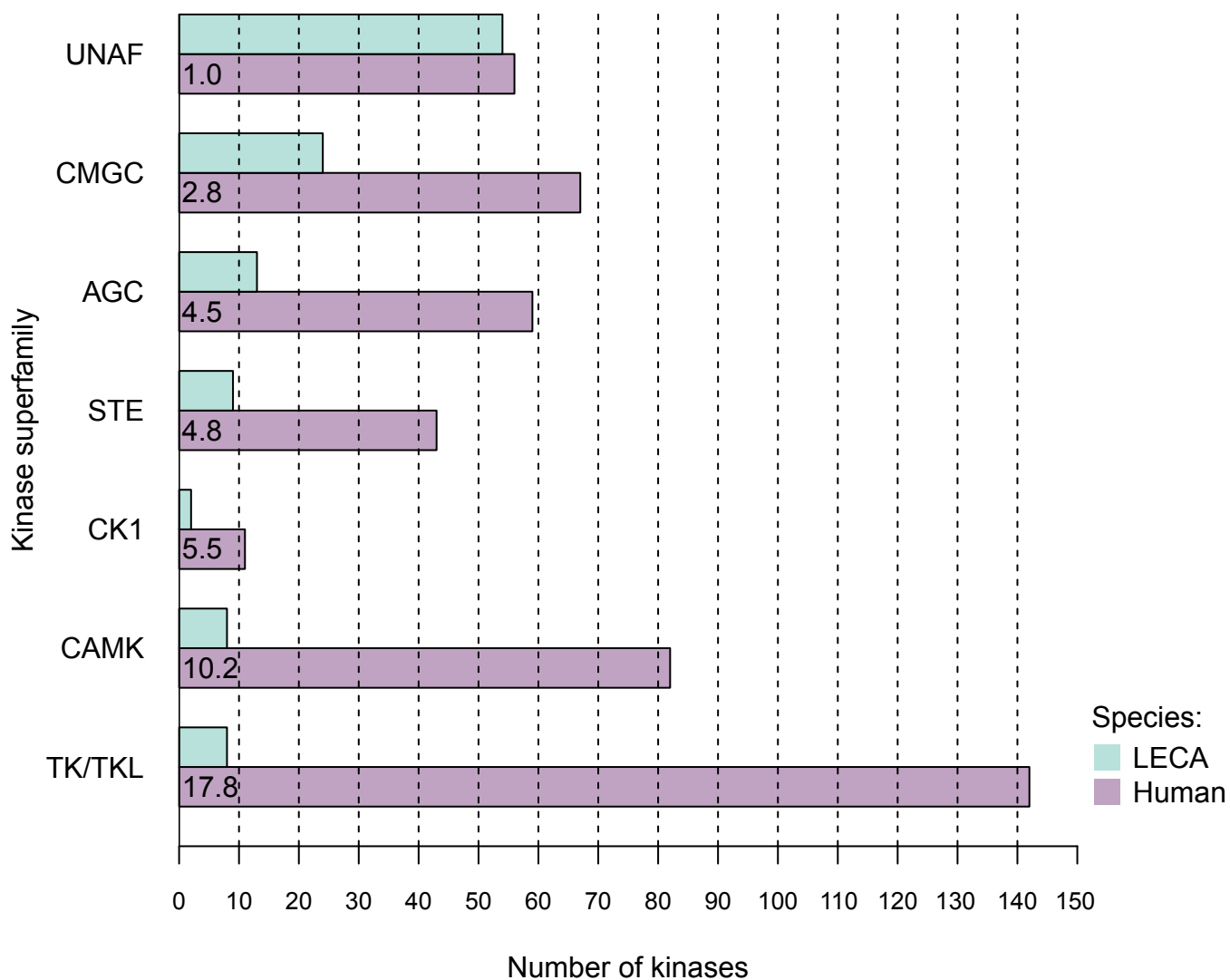


Fig. 5: The expansion of ePK superfamilies from LECA till human. Human bars show the multiplication factor between the number of kinases in the LECA and in human.

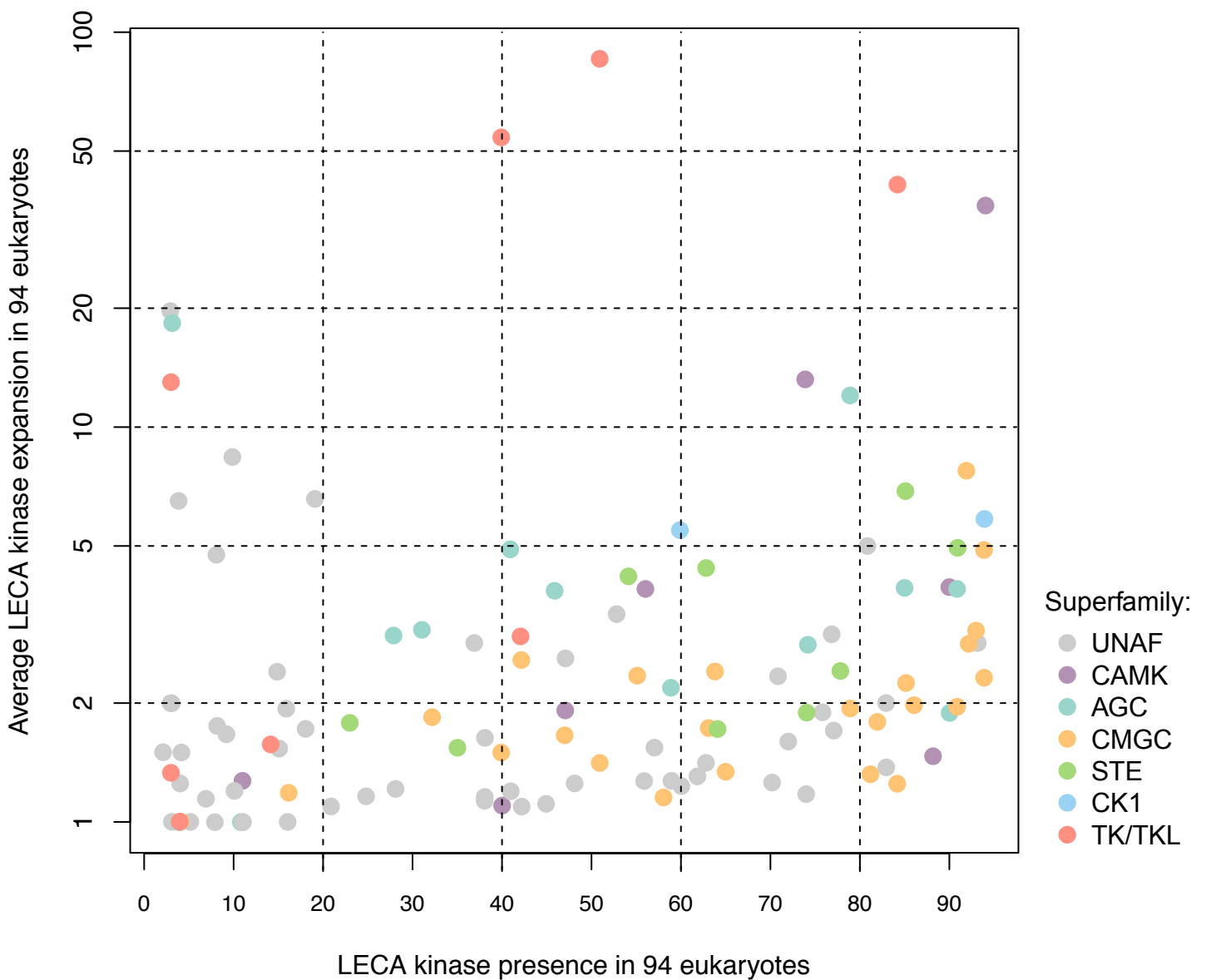


Fig. 6: EPK superfamily dynamics. The presence of 118 LECA kinases in 94 present-day eukaryotes is shown versus the average expansion of these LECA kinases in the same set of species. LECA kinases are colour coded according to ePK superfamily.