The origin, evolution and molecular diversity of the chemokine system

Alessandra Aleotti[1,2] *, Matthew Goulty[1,2] *, Clifton Lewis[1,2], Flaviano Giorgini[1,2] and Roberto Feuda[1,2] #

*Equal contribution

#Author for correspondence: rf190@leicester.ac.uk

[1] Department of Genetics and Genome Biology University of Leicester, Leicester, United Kingdom

[2]Neurogenetics Group, College of Life Sciences, University of Leicester, Leicester, United Kingdom

Keywords: Chemokine, Phylogeny, Evolution, Vertebrate, GPCR, Receptor, Ligand, CK, CKL, CKR, CKLF, TAFA, CYTL

ABSTRACT

Chemokine signalling performs key functions in cell migration via chemoattraction, such as attracting leukocytes to the site of infection during host defence. The system consists of a ligand, the chemokine, usually secreted outside the cell, and a chemokine receptor on the surface of a target cell that recognises the ligand. A number of non-canonical components interact with the system. While canonical components have been described in vertebrate lineages, the distribution of the non-canonical components is less clear. Uncertainty over the relationships between canonical and non-canonical components hampers our understanding of the evolution of the system. We used phylogenetic methods, including gene-tree to species-tree reconciliation to untangle the relationships between canonical and non-canonical components, identify gene duplication events and clarify the origin of the system. We found that unrelated ligand groups independently evolved chemokine-like functions. We found non-canonical ligands outside vertebrates, such as TAFA "chemokines" found in urochordates. In contrast, all receptor groups are vertebrate-specific and all - except ACKR1 - originated from a common ancestor in early vertebrates. Both ligand and receptor copy numbers expanded through gene duplication events at the base of jawed vertebrates, with subsequent waves of innovation occurring in bony fish and mammals.

INTRODUCTION

The chemokine system is responsible for regulating many biological processes, including host defence, neuronal communication and homeostasis (1–3). It has two components, a ligand, usually a small cytokine called a chemokine, and a receptor. The ligands are categorized into four groups, XC, CC, CXC, and CX3C, according to the pattern of cysteine residues in the N-terminal portion of the protein (4). Likewise, the receptors are classified based on the ligands they bind to into four groups, the XCR, CCR, CXCR, and CX3CR, and all of them belong to the GPCR class A superfamily (5). In addition to canonical components, other molecules have been discovered to function similarly to chemokine ligands (6) or receptors (7) (see Table S1). These include: the chemokine-like factor (CKLF) that binds to chemokine receptor CCR4 (8, 9) and drives cell migration *in vivo* (10); TAFA chemokines, expressed mainly in the nervous system, which share structural similarities to canonical chemokines (11, 12) and bind GPCRs related to chemokine receptors, e.g. formyl peptide receptors (13, 14) and GPR1 (15); Cytokine-like 1 (CYTL1) that binds CCR2 (16) and has been suggested to be related to CC ligands based on the presence of a IL8-like chemokine fold (17). There are also non-canonical chemokine receptors, such as: the chemokine-like receptor (CML1, or also CMKLR1) (18); atypical chemokine receptors (ACKRs) (19); and viral chemokine receptors (20–23). Unlike other chemokine receptors, atypical receptors cannot initiate classical chemokine signaling upon ligand binding (19, 24). The human genome encodes four types of atypical chemokine receptors: the ACKR1 (also known as DARC), ACKR2 (also known as D6), ACKR3 (also known as CXCR7) and ACKR4 (also known as CCRL1) (25, 26). Additionally, several proteins of viral origins, such as US28 from human cytomegalovirus, have chemokine-receptor/binding activity (20, 21). These viral proteins can bind a wide array of chemokine ligands (21).

Despite the extensive research on the chemokine system, with over 320,000 papers available on PubMed, many aspects of its evolution remain unclear. For instance, the homology between canonical and non-canonical ligands is uncertain and supported by circumstantial evidence, such as shared specific motifs (10, 11, 17, 27). Furthermore, the relationships between canonical,

atypical, and viral receptors and the outgroup of the canonical chemokine receptors remain uncertain. Finally, the evolutionary history of the canonical and non-canonical components remains poorly understood outside a few key model systems (5, 28, 29). These outstanding questions share common underlying causes, including the use of inadequate inference methods (such as relying solely on sequence similarities) and limited sampling of species (e.g., focusing mainly on humans, mice, and zebrafish (30, 31)). Additionally, solving the phylogenetic relationships for short molecules such as chemokine receptors and ligands is particularly challenging due to the lack of strong phylogenetic signals (32).

Here, to clarify these outstanding questions, we use state-of-the-art phylogenetic methods, including those designed for single-gene phylogenies, a large taxonomical sampling comprising both vertebrate and invertebrate genomes and the entire complement of canonical and non-canonical components of both receptors and ligands. Our findings substantially clarify the phylogenetic relationship between canonical and non-canonical ligands and receptors and suggest that unrelated proteins evolved "chemokine-like" ligand function multiple times independently. Additionally, we discovered that all the canonical and non-canonical chemokine receptors (except ACKR1) originated from a single duplication in the vertebrate stem group, which also gave rise to many GPCRs. Lastly, we characterized the complement of canonical and non-canonical components in the common ancestor of vertebrates and identified several other ligands and receptors with potential chemokine-related properties that could be explored in future functional work.

RESULTS

There are five unrelated groups of ligands.

Initially, we focused on the ligands, including all the canonical chemokines, the CYTL, the TAFAs and the CKLF Super Family (CKLFSF) proteins. The presence of a four transmembrane MARVEL domain in the latter proteins (10, 33, 34) distinguishes them from canonical chemokines, the CYTL and the TAFAs. Therefore, we separated these two groups for further analysis. Using BLASTP or PSI-

BLAST (35–37) (see Materials and Methods for more details) against 64 species from 19 animal phyla (Table S2), we identified 891 putative homologs for chemokines, TAFA and CYTL and 602 putative homologs of the CKLF Super Family.

We utilized CLANS (38, 39), a clustering tool based on sequence similarity and local alignment, to identify homology within these two groups. Unlike traditional phylogenetic methods, CLANS assigns homology between sequences based on BLAST and customizable stringency levels defined according to p-values (38). When two (or more) sequences are connected at a lower p-value (closer to 0), this indicates a high level of homology. Conversely, if two or more sequences only connect at a higher p-value, this suggests a relatively low level of sequence homology. Our analysis shows that canonical chemokines form a distinct group with a clear distinction between C-X-C-type and C-C-type (Figure 1A). Whereas, CXCL17, TAFA and CYTL remain separate from canonical chemokines and from each other even at the loosest p-values tested (Figure 1A). The distinction between CXCL17 and all other canonical chemokines is consistent with our receptor results showing that the receptor for CXCL17, GPR35 (40), is also not within the canonical chemokine receptor group (see below). Within the CKLFSF, two large clusters were identified, named CKLF I and CKLF II, although these ultimately connect to form one large superfamily (Figure 1C). These clusters are robust to the different stringency thresholds used (Figures S1 and S2 and Materials and Methods for further details). Our results indicate that even when the stringency level to detect homology is relaxed, canonical chemokines, TAFA, CYTL, and CXCL17 remain in distinct clusters. This suggests that, similarly to CKLFs, these proteins are not homologous and convergently evolved chemokine-like properties. We have thus identified five distinct groups of ligands: i) the canonical chemokines, ii) TAFA "chemokines", iii) CYTL, iv) CXCL17, and v) CKLF Super Family (Figure 1A and 1C).

The evolution of chemokine and chemokine-like ligands in animals.

To better understand the evolution of both canonical and non-canonical chemokine ligands, we performed a separate phylogenetic reconstruction for each group (Figure 1B, D and Figures S8-17) (see methods for details). To evaluate the nodal support, in addition to the UltraFast bootstrap

112 (UFB) (41, 42), we used Transfer Bootstrap Expectation (TBE), a method that has been developed for
113 single-gene phylogeny (43). To evaluate ortholog/paralog relationships and overall dynamics of the
114 ligand complement, we used GeneRax (44). This new method uses maximum likelihood to reconcile
115 the gene tree with the species tree (44). In brief, given a gene and species tree, GeneRax uses a
116 maximum likelihood approach to optimize the duplication and loss events (44, 45).

117 Our analysis initially identified a few invertebrate putative chemokine ligands (Figure 1A),
118 however, these sequences lacked protein signatures associated with the canonical ligands (Figures
119 S3-5 and File S3), and they were therefore excluded from further analysis (see Supplementary
120 Results for further information). The phylogenetic tree for the canonical ligands identifies two major
121 groups, the CC-type, which also includes the XC- and X3C-types, and the CXC-type (TBE = 0.95, UFB =
122 92%) (Figure 1B and Figures S8,9), confirming the previous finding obtained using synteny data (46,
123 47). Next, to clarify the distribution of canonical chemokines, we first reconciled their gene tree with
124 the species tree and then used the reconciled tree to trace the presence/absence of each chemokine
125 group throughout all the species (Figure 2A and Figure S18). Our results confirm previous findings
126 that canonical chemokines are uniquely present in vertebrates (28, 46). Additionally, they indicate
127 that chemokines are not evenly distributed across vertebrates. Some are very ancient, e.g., CXCL12
128 is present in lamprey; CXCL14 and CCL20 are present in all jawed vertebrates; and CXCL8 is present
129 in all jawed vertebrates except in the cartilaginous fishes lineage. However, a large part of the
130 chemokine diversity evolved within mammals (e.g., CXCL1/2/3, CXCL16, and CCL25), particularly
131 placentals (e.g., CXCL5/6 and CCL3/18). The complement of the canonical chemokines undergoes the
132 largest expansion at the base of jawed vertebrates, where there is an expansion from 4 to 18 genes
133 (Figure 2B). A second expansion occurred at the base of bony fishes (i.e., Osteichthyes) followed by a
134 relative stability until placental mammals, where the total number of canonical chemokine ligands
135 jumped to 45 genes. Finally, unlike previous works (48) our results support the presence of orthologs
136 of both CC-type and CXC-type in the common ancestor of all vertebrates (Figure 2A).

137 Differently from the canonical chemokines, we identified a *bona fide* TAFA, i.e., with specific
138 protein motifs, in the urochordates, the sister group to vertebrates (see Supplementary Results and
139 Figures S6-7). The phylogenetic trees (Supplementary Figures S10,11) identified monophyletic
140 groups for TAFA5 (TBE=0.98, UFB=98%), TAFA1 (TBE=0.94, UFB=98%), TAFA4 (TBE=0.77, UFB=75%)
141 and TAFA2/3 (TBE=0.65, UFB=84%). The reconciled tree from GeneRax places the root at the
142 urochordate sequence (Figure S19), therefore clarifying that the TAFA5 clade is the sister group to
143 TAFA1-4 (Figure 2A). The family originated in the ancestor of urochordates and vertebrates, and the
144 first duplications occurred at the base of vertebrates giving rise to the TAFA5 split followed by the
145 TAFA1 split. Subsequently, at the base of jawed vertebrates, additional duplications bring the
146 complements from 3 to 10 (Figure 2B), giving rise to the remaining groups so that all jawed
147 vertebrates possess the full diversity of TAFAs.

148 The phylogenetic trees for CYTL and CXCL17 mainly reflect the species trees (Figures S12-15),
149 and the reconciliations revealed very simple complement dynamics (Figure 2B and Figures S20,21).
150 However, these molecules show a remarkable difference in their distribution. CYTLs are present
151 throughout gnathostomes, while CXCL17 is found only in placental mammals (Figure 2A).

152 The phylogenetic analysis for the CKLF super family (Figure 1D and Figures S16,17) recovered
153 a monophyletic clade for the CKLF I group (TBE=0.96, UFB=80%) that we had already identified
154 through CLANS. This group contains CKLF, that is known to interact with C-C chemokine receptor 4
155 (8, 9), as well as CMTM1, 2, 3, 5, and proteolipid protein 2 (PLP2). Other monophyletic clades that
156 are consistent with the CLANS are CMTM4/6 (TBE=0.90, UFB=61%), CMTM7 (TBE=0.92, UFB=83%)
157 and a clade containing CMTM8 plus other related molecules such as plasmolipin (PLLP) and myelin
158 and lymphocyte proteins (MAL) (TBE=0.89, UFB=60%). The latter were all part of a large cluster that
159 we called CKLF II in the CLANS (Figure 1C). However, the placement of the root of the tree in Figure
160 1D can affect the interpretation of the relationships among CKLF II subgroups. To address this
161 problem and clarify the patterns of duplications and the presence/absence of each group
162 throughout animals, we used GeneRax to reconcile the gene with the species tree (see above and

163 Material and Methods for details). Our results suggest (Figure 2 and Figure S22) that most CKLFSF
164 groups, such as CMTM4,6 and 8, originate in the vertebrate stem group from pre-existing CMTM
165 genes and are widely distributed in animals. The CKLF I subgroups originate from duplications at the
166 base of jawed vertebrates, except for the split between CKLF and CMTM1 that occurs only within
167 mammals (Figure 2A). We observe the major two expansions of the CKLFSF genes in the stem group
168 of vertebrates (from 6 to 10 complements), and then in jawed vertebrates (from 10 to 16
169 complements). Interestingly the extents of these expansions are less drastic than those we see for
170 canonical chemokines (Figure 2B). In total, we have identified that the five distinct ligand groups
171 have a different origin in the animal tree of life and underwent divergent evolutionary histories.

172 **Canonical and non-canonical chemokine receptors are divided into four groups.**

173 Next, we investigated the origin and pattern of duplication for the chemokine and
174 chemokine-like receptors. Using BLASTP against the 64 species, we identified 7,157 putative
175 chemokine receptors (see Materials and Methods for more details), and we investigated their
176 relationships using CLANS (see above for justification). The result (Figure S23C) identifies four main
177 groups of chemokine and chemokine-like receptors. The first comprises canonical receptors (i.e.,
178 CCR, CXCR, CX3CR1, CX3C, and XCR1), and the second includes atypical receptor 3 and GPCR182,
179 which has been recently shown to have chemokine activity (49). The third group, which we named
180 Chemokine-like plus (CML-plus), contains the chemokine-like receptors (CML1 also known as
181 chemerin receptor 1), formyl peptide receptors (FPR) that bind the TAFA ligands (13, 14) and other
182 GPCRs such as GPCR1 (chemerin receptor 2), GPCR33, PTGDR2. Furthermore, the CLANS analysis
183 identifies an intermediate group containing angiotensin, apelin and other receptors and shows
184 sequence similarity to canonical and chemokine-like receptors (Figure S23B). Finally, our analysis
185 identifies a small cluster composed of only ACKR1 that do not connect to other GPCRs or other
186 atypical receptors even at loose p-value thresholds. This indicates that their sequence is either non-
187 homologous or highly divergent from other chemokines and atypical receptors. Overall, these
188 groups are robust to the stringency threshold used (i.e., different p-values) (Figure S23).

Interestingly, no specific cluster of viral or viral-like receptors was identified, but 6 of the reference viral receptor sequences clustered with the canonical chemokine receptors.

Altogether, these results confirm the homology between the canonical receptors and atypical receptor 3/GPCR182. However, the results indicate that the other GPCRs, such as the chemokine-like receptors, formyl peptide receptors, GPCR1, and GPCR33, are also closely related to the canonical receptors. Remarkably, these results also indicate that ACKR1 is not homologous to the canonical chemokine receptors. Furthermore, all clusters of chemokine receptors contained only vertebrate sequences, except for the receptors of viral origin.

Canonical and chemokine-like receptors derive from single gene duplication in the ancestor of vertebrates.

Previous studies suggested that the chemokine receptors evolved from a duplication of angiotensin receptors (50) or adrenomedullin receptors (28, 51). However, these works were based on error-prone phylogenetic methods such as Neighbour Joining (51). Our CLANS results indicate that chemokine receptors and chemokine-like receptors have only been observed in vertebrates. Therefore, we need to focus on invertebrate genomes to clarify the chemokine's outgroup. To clarify this, we lowered the p-value thresholds of CLANS (to p-value < $1e^{-50}$) and collected a combined dataset including all chemokine sequences and outgroups (i.e., sequences that connect to the chemokine cluster), resulting in 3,026 sequences. We then performed a phylogenetic tree on this dataset using maximum likelihood methods with UFB and TBE for evaluating nodal support (see above and Materials and Methods for details).

Our combined phylogenetic analysis shows strong support for the monophyly of canonical chemokines (UFB=96, TBE=1.0), the CML-plus (UFB=95, TBE=0.99) and the atypical 3/GPCR182 (UFB=100, TBE=1) (Figure 3, S24 and S25). In contrast, viral chemokines are paraphyletic, with three sequences placed within the canonical chemokines and 3 forming a monophyletic group sister to them (UBF=84 TBE=1.0). Our results also suggest that the intermediate group, which includes apelin receptors, angiotensin receptors, bradykinin receptors, and orphan GPCRs (e.g., GPCR25; GPCR15)

215 forms a monophyletic clade with the canonical chemokines, CML-plus group and atypical3/GPCR182
216 (UFB=61, TBE=0.91). However, its position changes between the sister group to canonical
217 chemokines plus atypical3/GPCR182 in the TBE tree (TBE=0.84) and sister to CML plus in the ultrafast
218 bootstrap tree (UFB=38).

219 All the groups mentioned above form a large clade composed of vertebrate-specific GPCRs
220 (UBF=100 TBE=0.96) that also includes other GPCRs, such as CLTR and P2RY receptors (Figure 3, S24
221 and S25). Another orphan GPCR, GPR35, had been proposed as a potential chemokine receptor
222 (putative CXCR8) (40). Our analysis collected GPR35 and placed it within this large vertebrate specific
223 clade indicating that is it also a vertebrate specific gene but not phylogenetically a 'canonical'
224 chemokine. The closest outgroup to this clade is composed of a few sequences from urochordates,
225 the sister group of vertebrates (UFB=49 TBE=0.91) (Figure 3, S24 and S25). Interestingly, as the sister
226 group of this clade, we identify a group composed of Relaxin receptors which contain sequences
227 from both urochordates and vertebrates (UBF=53 TBE=0.95). Finally, as the sister group of these
228 large clades, we identified a clade of cephalochordate-specific sequences (UBF=44).

229 To clarify the duplication pattern and origin of the chemokine receptors, we used GeneRax
230 (44) (see Materials and Methods). Our results indicate (Figure 4A, S26) that all chemokine receptors
231 (except ACKR1) originated from a duplication in the stem lineage of vertebrates. This duplication of
232 an unknown GPCR gave rise to the CML-plus, the canonical chemokines and atypical 3/GPCR182
233 groups as well as the intermediate group and other GPCRs (Figure 4A, S26). This result is consistent
234 with the distribution of the paralogous Relaxin receptors which are present both in urochordates
235 and vertebrates and the position of the orphan urochordate sequences as sister group of canonical
236 chemokines, CML-plus group and atypical3/GPCR182 and other GPCRs (see above).

237 We used the reconciliation to better understand the repertoires of receptors present at key
238 nodes during vertebrate evolution. Our results (Figure 4B) show a substantial difference in the
239 duplication pattern of different receptor families. For example, the complement of the
240 atypical3/GPCR182 remains constant throughout vertebrate evolution while the canonical and

chemokine-like groups expanded dramatically. The canonical chemokine expanded from 5 to 20 genes and the CML-plus from 1 to 11 in the ancestor of the jawed vertebrates (Figure 4B). The expansion of the canonical CKRs is also not evenly distributed across its subgroups, with the ancestral CC type receptors undergoing a series of duplications in jawed vertebrates while the CXCR paralogs did not, specifically one (CXCR4) remains in single copy across all vertebrates. We inferred that in the stem lineage of vertebrates, five canonical chemokine paralogs had already diverged, representing the two major types of receptors (2 CCR and 3 CXCR paralogs). Also, present in the stem lineage of vertebrates were ACKR3 and GPCR182 as well as a single copy gene which would later diverge to produce all the CML-plus clade.

DISCUSSION

This work substantially clarifies the evolutionary assembly of the chemokine system. Our analysis shows that, contrary to the receptors which evolved from a single duplication event in the vertebrate stem group, several unrelated molecules acquired the ability to interact with chemokine receptors over the course of evolutionary history. Furthermore, our results (summarized in Figure 5) suggest that the key components of the chemokine system, including the chemokine receptors themselves, evolved in the stem group of vertebrates in the Ediacaran/Cambrian boundary around 540 million years ago and then underwent substantial diversification in the stem group of jawed vertebrates. These findings shed new light on the complex evolutionary history of the chemokine system.

Unrelated molecules converged to chemokine function.

Based on the presence of shared protein motifs, TAFA “chemokines” (11, 12), CXCL17 (52, 53) and CYTL (17) have been proposed to be homologous to chemokine ligands. However, our findings strongly suggest that these molecules are not homologous (Figure 1) and likely acquired the ability to activate a chemokine-like response through convergent evolution. Our conclusions differ from those previous studies (11, 17, 52, 53) due to the differences in data completeness and

methodological approach. Specifically, we used a complete set of canonical and non-canonical ligands, and assessed the homology using overall sequence similarity rather than single motifs. Our results support and expand upon the findings of (27), which suggested that the presence of a CXC or CC motif is necessary but not sufficient for a protein to be defined as a chemokine ligand. Similarly, CKLF has been considered a "new member" of the chemokine family based on its function (10) we argue that classification based solely on function is insufficient and can be misleading. Instead, we recommend considering the evolutionary relationships among these molecules as the primary criterion for classification.

Most of the canonical and non-canonical ligands are vertebrate innovations.

Our results clarify the distribution of canonical chemokine ligands in animals (Figure 2) and confirm that they are present only in vertebrates (28). We identify orthologs of CXCL and CCL ligands in both extant lineages of cyclostomes (Figure 2A). While chemokines have already been described in lamprey (48, 54, 55), it is the first time, to our knowledge, that they are described also in hagfish. Our findings also indicate that both CC and CXC types were present in the common ancestor of all vertebrates and that few ancestral genes gave rise to the entire diversity of ligands that we know in current animals. Furthermore, our results indicate that many chemokines, such as CXCL1-7, CXCL16, as well as CCL25, CCL11/13, and CCL2/7, are uniquely present in mammals suggesting that the mammal ligand repertoire is substantially more complex than the one observed in other vertebrates.

Regarding non-canonical chemokine-like families, our findings indicate that the TAFA family originated in the ancestor of vertebrates and urochordates; CYTL is a novelty of jawed vertebrates; and CXCL17 is mammal-specific and likely unrelated to canonical chemokines (similar to its receptor, GPR35 (40), that is not a canonical chemokine receptor). The CKLF super family has a more complex pattern with the presence of few groups in invertebrates and then great expansions occurring at the base of vertebrates. The CKLFSF includes a monophyletic clade (CKLF group I) comprising the original CKLF that binds CCR4, as well as CMTM1,2,3,5, derived from duplications at the jawed vertebrates stem group. Interestingly, our analysis also revealed that additional molecules not previously

293 considered part of the CKLF super family are closely related to classic members and should be
294 included in it. For example, proteolipid protein 2 (PLP2) belongs to the CKLF I group and is, therefore
295 more closely related to the CKLF with chemokine function than several other CKLFSF members.
296 Similarly, CMTM8 is more closely related to plasmolipin (PLLP) and myelin and lymphocyte protein
297 (MAL) than to any of the classic CKLFSF members. Although this relationship had been proposed
298 based only on sequence similarity (33), our phylogenetic analysis provides additional evidence for it.
299 Therefore, all these additional members should also be explored for potential chemokine-like
300 functions in both vertebrates and invertebrates.

301 **All but one receptor derive from a single gene duplication.**

302 Our results clarify the distribution of canonical chemokine receptors in vertebrates (Figure
303 4), and their evolutionary relationships and identify the pattern of duplication that leads to their
304 origin (Figure 4A, S26). Unlike previous works (56), we identify that atypical receptors do not form a
305 monophyletic group. Specifically, atypical 2 and 4 are part of the canonical clade specifically related
306 to CC-type receptor subclades. Furthermore, we find that the atypical 3 receptors are related to
307 GPCR182, supporting previous functional data suggesting that the latter are atypical chemokine
308 receptors binding CXCL10, 12, and 13 (49). We attribute these differences to our use of wider GPCR
309 sampling and improved methods for phylogenetic inference.

310 Remarkably, our results do not identify ACKR1 as related to the main chemokine receptors
311 but rather as a divergent clade (Figure S23). To our knowledge, this is the first time this observation
312 has been made. Our current results do not allow us to clarify the evolutionary origin ofACKR1.
313 However, the presence of 7TMD domains suggests that they are GPCRs that independently acquired
314 the ability to bind chemokines. Alternatively, similarly to other genes evolved in the immune system,
315 ACKR1 may have been subjected to strong selective pressures that substantially changed their
316 sequence, obscuring their phylogenetic relationships.

317 Viral chemokines represent a cryptic group that can bind multiple chemokines (20, 21).
318 Despite their functional similarity to canonical chemokines, viral chemokines' evolutionary origin and

distribution remain poorly understood. Our results indicate that viral GPCRs do not form a monophyletic group, suggesting that the ability to encode chemokine-like receptors has evolved independently in multiple viruses, including cytomegaloviruses and poxviruses. However, further research is needed to clarify the distribution and origins of viral GPCRs.

Our analysis reveals that the clade comprising apelin receptors, angiotensin receptors, bradykinin receptors, and orphan GPCRs (shown in Figure 3, 4 and S24–26) is closely related to chemokine receptors. This finding partially supports previous studies (50) that suggested a gene duplication event gave rise to both chemokines and angiotensin receptors. Interestingly, we found that single gene duplication in the vertebrate stem group led to the emergence of canonical receptors and atypical 2,3,4, GPCR182, chemokine-like, formyl peptide receptors, the intermediate group, and many other known and orphan GPCRs including putative chemokine receptor GPR35. These findings suggest that two rounds of genome duplication (57, 58) played a role in the expansion of GPCR gene families. Future research will focus on investigating the functions of the orphan genes and many-to-one orthologs discovered in urochordates. This will provide further insight into the evolution and diversification of GPCR families in vertebrates.

The molecular assembly of the chemokine system

In this work, we explored the evolution of both ligand and receptor components of the chemokine signaling system, including non-canonical molecules with either chemokine-like function or sequence similarity. Our analysis suggests that the canonical chemokine signaling evolved in the vertebrate stem group (about 540 Mya) likely due to the two rounds of genome duplication that gave rise to many vertebrate novelties (57, 58). We found that the ancestral vertebrate repertoire included orthologs of both major ligand groups (CXCL and CCL) and both CCR and CXCR receptors and non-canonical components such as TAFA and CKLFSF ligands, and the receptors Atypical 3 and GPCR182 (Figure 5). The distribution of ligands and receptors in the ancestor of all vertebrates, seems to confirm the hypothesis that the ancestral function of chemokines was homeostatic (e.g.,

CXCL12, CXCL14) with inflammatory functions arising from recent duplications (e.g., CXCL5, CXCL6), potentially reflecting a rapid evolution induced by the selective pressure of new pathogens (30).

The evolutionary history of canonical components includes several examples of known ligand-receptor pairs following a corresponding pattern of origin and temporal dynamics of duplications. This is true for example, for the ancient homeostatic CXCL12 ligand and its corresponding receptors CXCR4 and ACKR3, that all originated in early vertebrates (30). The early origin and conservation of CXCR4 and CXCL12 in the ancestor of vertebrates is interesting as this pair plays a key role in the migration of neural crest cells (59) - a key vertebrate innovation (60). For many of the non-canonical components, however, the ligand-receptor interactions are largely unclear, and their pattern throughout vertebrate evolution remains to be explored. Overall, our results indicate that three waves of molecular innovation in the vertebrates, jawed vertebrates, bony fishes and mammal stem group increased the chemokine system's molecular complexity (Figure 5), allowing for the fine-tuning present in modern-day animals.

MATERIALS AND METHODS

Data Mining and Dataset Assembly. We collected 64 proteomes from 25 vertebrates, six chordates and 33 other animals covering the whole animal tree (Table S2). BUSCO v4.0.6 (61, 62) and the metazoa_odb10 set of 954 genes were used to evaluate their completeness (Table S2).

To identify potential homologs of canonical chemokines, TAFA chemokines and CYTL1, we used 207 curated sequences that we obtained from SwissProt (63, 64) as seeds for an initial BLASTP (35, 37) with e-value < e^{-10}. To identify putative chemokines in cyclostomes, the lamprey *Petromyzon marinus* (65) and the hagfish *Eptatretus burger* (66), we loosened the e-value to 0.05. Where putative chemokine sequences were found for one cyclostome species but not the other, those sequences were used to search again the other species. Furthermore, to investigate the presence of ligands outside vertebrates, we performed an additional BLASTP on invertebrate proteomes with an even looser e-value (0.1) and collected only up to five hits. This provided 18 initial candidate

370 homologs spanning multiple invertebrate phyla. Further characterisation of these invertebrate
371 sequences, through BLASTP versus SwissProt, protein domains search with InterProScan (67, 68),
372 position in CLANS analysis (see below) and, where necessary, multiple sequence alignments, lead us
373 to retain only one urochordate sequence as a putative TAFA homolog (see Supplementary Results
374 for details).

375 To identify homologs for the CKLF superfamily, we used 21 SwissProt-reviewed sequences.
376 In addition to the BLASTP search, we used a position-specific iterative BLAST (PSI-BLAST) (36) with an
377 e-value threshold of $<e^{-10}$. Using this approach, we identified a total of 590 putative homologs,
378 including 186 from invertebrates.

379 We used BLASTP using 178 manually annotated receptor sequences from SwissProt as query
380 sequences for the chemokine receptors. This includes all human canonical and atypical chemokine
381 receptors (69). We also collected 8 viral sequences with chemokine receptor activity from UniProt
382 (70) and performed a second BLASTP. We extracted all BLAST hits with e-values $< e^{-10}$ and used
383 Phobius (71) to predict their transmembrane domain structure. Only sequences with 5-8
384 transmembrane domains were kept. Hit sequences were annotated by their top 5 BLAST hits against
385 SwissProt. All hits from both BLASTs were merged and filtered by cd-hit (72, 73) to remove
386 redundant sequences at the 95% similarity threshold. This resulted in 7,157 putative chemokine
387 GPCR sequences.

388

389 **Identification of subgroups with Cluster Analysis of Sequences (CLANS).** We utilized CLANS (38, 39)
390 with default parameters and different p-values (i.e., stringency values) to visualize the relationships
391 between subgroups of ligands and receptors. We assessed the similarity and interrelationships
392 between different clusters by gradually relaxing the p-value threshold (Figures S1, S2 and S23).
393 Additionally, we annotated each cluster using gene annotations for key species *Homo*, *Mus*, *Gorilla*,
394 *Gallus, Anolis* and *Danio*. In the case of the receptors, to improve the cluster annotation all human

Class-A GPCRs (excluding olfactory receptors) from GPCRdb (74) were added to the dataset as well as the 8 seed viral chemokine receptors from UniProt (70).

Alignment and phylogenetic analysis

Alignment. All ligand and receptor sequences were aligned using MAFFT (75, 76) with the --auto setting and using trimAl (77) to remove positions with >70% gaps.

Gene Trees. All gene alignments were analysed using IQTREE2 (78), the model test algorithm (79) was used to select the best substitution model for each analysis. Best models selected by IQTREE2 for each set are listed in Table S3 (for receptors we manually selected GTR20+F+G4 as the model as it was a large dataset). Nodal support was estimated using 1,000 ultrafast bootstrap (41, 42) replicates. All analyses were repeated to run 100 non-parametric bootstrap repeats to calculate nodal support with transferable bootstrap expectation: which is specifically designed to account for phylogenetic instability (43).

For the receptors, due to the high computational burden of running TBE analyses on sequence-dense datasets, we first analysed the full set of 3,026 sequences connected in CLANS at a p-value of $< 1\ e^{-50}$ using UFB (Figure S25). Then, we extracted the chordate-specific clade sequences, including all chemokine groups and their immediate outgroups, to analyse using TBE.

Gene tree species tree reconciliation. To understand the pattern of duplication and the evolution of gene complement we used GeneRax (44). GeneRax requires a gene tree that was obtained as described above and a species tree that we constructed manually using publicly available information. In the instances where the genes tree contained polytomies, we used ETE3 (80) to solve them. The undated DL mode and the closest approximation of the best-fitting substitution model were used for each alignment. To track the evolution of sub-lineages within each group, we used annotated sequences of key species (e.g., *Homo sapiens* and *Mus musculus*) as reference. For the receptors, we used only the chordate-specific clade subtree and sequences due to the

computational burden of running GeneRax on a high number of sequences. For species tree-gene tree reconciliation, we treat the viral sequences as human sequences.

DATA AVAILABILITY

Supplementary material and raw output files for all the analyses described in this paper are available in the GitHub page: Roberto-Feuda-Lab/Chemokine2023 (github.com).

ACKNOWLEDGEMENTS

This work is supported by a University Research Fellowship (UF160226) to RF. AA is supported by a Research Grant from the Royal Society to RF (RGF\R1\181012). MG is supported by a PhD Scholarship from the University of Leicester. CL is supported by a BBRSC MIBPT fellowship. This research used the ALICE High-Performance Computing Facility at the University of Leicester.

REFERENCES

1. X. Blanchet, M. Langer, C. Weber, R. Koenen, P. von Hundelshausen, Touch of Chemokines. *Frontiers in Immunology* **3**, 175 (2012).
2. P. López-Cotarelo, C. Gómez-Moreira, O. Criado-García, L. Sánchez, J. L. Rodríguez-Fernández, Beyond Chemoattraction: Multifunctionality of Chemokine Receptors in Leukocytes. *Trends in Immunology* **38**, 927–941 (2017).
3. P. B. Tran, R. J. Miller, Chemokine receptors: signposts to brain development and disease. *Nature Reviews Neuroscience* **4**, 444–455 (2003).
4. A. Zlotnik, O. Yoshie, Chemokines: A New Classification System and Their Role in Immunity. *Immunity* **12**, 121–127 (2000).
5. H. Nomiyama, N. Osada, O. Yoshie, A family tree of vertebrate chemokine receptors for a unified nomenclature. *Developmental & Comparative Immunology* **35**, 705–715 (2011).

446 6. K. Zhang, S. Shi, W. Han, Research progress in cytokines with chemokine-like function.
447 *Cellular & Molecular Immunology* **15**, 660–662 (2018).
448 7. K. Chen, *et al.*, Chemokines in homeostasis and diseases. *Cell Mol Immunol* **15**, 324–334
449 (2018).
450 8. Y. Wang, *et al.*, Chemokine-like factor 1 is a functional ligand for CC chemokine receptor 4
451 (CCR4). *Life Sciences* **78**, 614–621 (2006).
452 9. Y. Wang, *et al.*, Two C-terminal peptides of human CKLF1 interact with the chemokine
453 receptor CCR4. *The International Journal of Biochemistry & Cell Biology* **40**, 909–919 (2008).
454 10. D.-D. Liu, *et al.*, Progress in pharmacological research of chemokine like factor 1 (CKLF1).
455 *Cytokine* **102**, 41–50 (2018).
456 11. Y. Tom Tang, *et al.*, TAFA: a novel secreted family with conserved cysteine residues and
457 restricted expression in the brain. *Genomics* **83**, 727–734 (2004).
458 12. D. C. Sarver, X. Lei, G. W. Wong, FAM19A (TAFA): An Emerging Family of Neurokines with
459 Diverse Functions in the Central and Peripheral Nervous System. *ACS Chem. Neurosci.* **12**, 945–958
460 (2021).
461 13. W. Wang, *et al.*, FAM19A4 is a novel cytokine ligand of formyl peptide receptor 1 (FPR1) and
462 is able to promote the migration and phagocytosis of macrophages. *Cell Mol Immunol* **12**, 615–624
463 (2015).
464 14. M. Y. Park, *et al.*, FAM19A5, a brain-specific chemokine, inhibits RANKL-induced osteoclast
465 formation through formyl peptide receptor 2. *Sci Rep* **7**, 15575 (2017).
466 15. C. Zheng, *et al.*, FAM19A1 is a new ligand for GPR1 that modulates neural stem-cell
467 proliferation and differentiation. *The FASEB Journal* **32**, 5874–5890 (2018).
468 16. X. Wang, *et al.*, Cytokine-like 1 Chemoattracts Monocytes/Macrophages via CCR2. *The*
469 *Journal of Immunology* **196**, 4090–4099 (2016).
470 17. A. Tomczak, M. T. Pisabarro, Identification of CCR2-binding features in Cytl1 by a CCL2-like
471 chemokine model. *Proteins: Structure, Function, and Bioinformatics* **79**, 1277–1292 (2011).

472 18. T. Yoshimura, J. J. Oppenheim, Chemokine-like receptor 1 (CMKLR1) and chemokine (C–C
473 motif) receptor-like 2 (CCRL2); Two multifunctional receptors with unusual properties. *Experimental*
474 *Cell Research* **317**, 674–684 (2011).
475 19. R. Bonecchi, G. J. Graham, Atypical Chemokine Receptors and Their Roles in the Resolution
476 of the Inflammatory Response. *Frontiers in Immunology* **7**, 224 (2016).
477 20. T. N. Kledal, M. M. Rosenkilde, T. W. Schwartz, Selective recognition of the membrane-
478 bound CX3C chemokine, fractalkine, by the human cytomegalovirus-encoded broad-spectrum
479 receptor US28. *FEBS Letters* **441**, 209–214 (1998).
480 21. T. F. Miles, *et al.*, Viral GPCR US28 can signal in response to chemokine agonists of nearly
481 unlimited structural degeneracy. *eLife* **7**, e35850 (2018).
482 22. M. M. Rosenkilde, M. J. Smit, M. Waldhoer, Structure, function and physiological
483 consequences of virally encoded chemokine seven transmembrane receptors. *British Journal of*
484 *Pharmacology* **153**, S154–S166 (2008).
485 23. H. Daiyasu, W. Nemoto, H. Toh, Evolutionary Analysis of Functional Divergence among
486 Chemokine Receptors, Decoy Receptors, and Viral Receptors. *Frontiers in Microbiology* **3** (2012).
487 24. M. Meyrath, *et al.*, The atypical chemokine receptor ACKR3/CXCR7 is a broad-spectrum
488 scavenger for opioid peptides. *Nature Communications* **11**, 3033 (2020).
489 25. R. J. B. Nibbs, G. J. Graham, Immune regulation by atypical chemokine receptors. *Nat Rev*
490 *Immunol* **13**, 815–829 (2013).
491 26. F. Bachelerie, *et al.*, New nomenclature for atypical chemokine receptors. *Nat Immunol* **15**,
492 207–208 (2014).
493 27. S. S. Denisov, CXCL17: The Black Sheep in the Chemokine Flock. *Frontiers in Immunology* **12**,
494 2811 (2021).
495 28. M. E. DeVries, *et al.*, Defining the Origins and Evolution of the Chemokine/Chemokine
496 Receptor System. *The Journal of Immunology* **176**, 401 (2006).

29. B. Bajoghli, Evolution and function of chemokine receptors in the immune system of lower vertebrates. *European Journal of Immunology* **43**, 1686–1692 (2013).

30. A. Zlotnik, O. Yoshie, The Chemokine Superfamily Revisited. *Immunity* **36**, 705–716 (2012).

31. H. Nomiyama, *et al.*, Extensive expansion and diversification of the chemokine gene family in zebrafish: Identification of a novel chemokine subfamily CX. *BMC Genomics* **9**, 222 (2008).

32. J. F. Fleming, R. Feuda, N. W. Roberts, D. Pisani, A Novel Approach to Investigate the Effect of Tree Reconstruction Artifacts in Single-Gene Analysis Clarifies Opsin Evolution in Nonbilaterian Metazoans. *Genome Biology and Evolution* **12**, 3906–3916 (2020).

33. W. Han, *et al.*, Identification of eight genes encoding chemokine-like factor superfamily members 1–8 (CKLFSF1–8) by in silico cloning and experimental validation. *Genomics* **81**, 609–617 (2003).

34. H.-J. Duan, X.-Y. Li, C. Liu, X.-L. Deng, Chemokine-like factor-like MARVEL transmembrane domain-containing family in autoimmune diseases. *Chinese Medical Journal* **133** (2020).

35. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

36. S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

37. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

38. T. Frickey, A. Lupas, CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).

39. F. Gabler, *et al.*, Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics* **72**, e108 (2020).

40. J. L. Maravillas-Montero, *et al.*, Cutting Edge: GPR35/CXCR8 Is the Receptor of the Mucosal Chemokine CXCL17. *The Journal of Immunology* **194**, 29–33 (2015).

41. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution* **30**, 1188–1195 (2013).

42. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).

43. F. Lemoine, *et al.*, Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).

44. B. Morel, A. M. Kozlov, A. Stamatakis, G. J. Szöllősi, GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution* **37**, 2763–2774 (2020).

45. T. A. Williams, *et al.*, The power and limitations of species tree-aware phylogenetics. 2023.03.17.533068 (2023).

46. H. Nomiyama, N. Osada, O. Yoshie, Systematic classification of vertebrate chemokines based on conserved synteny and evolutionary history. *Genes to Cells* **18**, 1–16 (2013).

47. A. Zlotnik, O. Yoshie, H. Nomiyama, The chemokine and chemokine receptor superfamilies and their molecular evolution. *Genome Biol* **7**, 243–243 (2006).

48. Z. Sun, *et al.*, The evolution and functional characterization of CXC chemokines and receptors in lamprey. *Developmental & Comparative Immunology* **116**, 103905 (2021).

49. A. Le Mercier, *et al.*, GPR182 is an endothelium-specific atypical chemokine receptor that maintains hematopoietic stem cell homeostasis. *Proceedings of the National Academy of Sciences* **118**, e2021596118 (2021).

50. P. Liò, M. Vannucci, Investigating the evolution and structure of chemokine receptors. *Gene* **317**, 29–37 (2003).

51. R. Fredriksson, M. C. Lagerström, L.-G. Lundin, H. B. Schiöth, The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol Pharmacol* **63**, 1256–1272 (2003).

52. M. T. Pisabarro, *et al.*, Cutting Edge: Novel Human Dendritic Cell- and Monocyte-Attracting Chemokine-Like Protein Identified by Fold Recognition Methods. *The Journal of Immunology* **176**, 2069–2073 (2006).

53. E. J. Weinstein, *et al.*, VCC-1, a novel chemokine, promotes tumor growth. *Biochemical and Biophysical Research Communications* **350**, 74–81 (2006).

54. A. M. Najakshin, L. V. Mechetina, B. Y. Alabyev, A. V. Taranin, Identification of an IL-8 homolog in lamprey (Lampetra fluviatilis): early evolutionary divergence of chemokines. *European Journal of Immunology* **29**, 375–382 (1999).

55. B. Bajoghli, *et al.*, Evolution of Genetic Networks Underlying the Emergence of Thymopoiesis in Vertebrates. *Cell* **138**, 186–197 (2009).

56. L. Pan, J. Lv, Z. Zhang, Y. Zhang, Adaptation and Constraint in the Atypical Chemokine Receptor Family in Mammals. *BioMed Research International* **2018**, 9065181 (2018).

57. M. Kasahara, The 2R hypothesis: an update. *Current Opinion in Immunology* **19**, 547–552 (2007).

58. O. Simakov, *et al.*, Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* **4**, 820–830 (2020).

59. W. Tang, Y. Li, A. Li, M. E. Bronner, Clonal analysis and dynamic imaging identify multipotency of individual Gallus gallus caudal hindbrain neural crest cells toward cardiac and enteric fates. *Nat Commun* **12**, 1894 (2021).

60. J. R. York, D. W. McCauley, The origin and evolution of vertebrate neural crest cells. *Open Biology* **10**, 190285 (2020).

61. R. M. Waterhouse, *et al.*, BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* **35**, 543–548 (2018).

62. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).

63. E. Boutet, *et al.*, "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View" in *Plant Bioinformatics: Methods and Protocols*, Methods in Molecular Biology., D. Edwards, Ed. (Springer, 2016), pp. 23–54.

64. S. Poux, *et al.*, On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* **33**, 3454–3460 (2017).

65. J. J. Smith, *et al.*, Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat Genet* **45**, 415–421 (2013).

66. D. Yu, *et al.*, Hagfish genome illuminates vertebrate whole genome duplications and their evolutionary consequences. 2023.04.08.536076 (2023).

67. E. M. Zdobnov, R. Apweiler, InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

68. P. Jones, *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

69. F. Bachelerie, *et al.*, Chemokine receptors (version 2020.5) in the IUPHAR/BPS Guide to Pharmacology Database. *IUPHAR/BPS Guide to Pharmacology CITE* **2020** (2020).

70. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).

71. L. Käll, A. Krogh, E. L. L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Research* **35**, W429–W432 (2007).

72. W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).

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

74. G. Pándy-Szekeres, *et al.*, GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Research* **51**, D395–D402 (2023).

75. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).

76. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

77. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

78. B. Q. Minh, *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

79. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).

80. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* **33**, 1635–1638 (2016).

81. D. H. Erwin, *et al.*, The Cambrian Conundrum: Early Divergence and Later Ecological Success in the Early History of Animals. *Science* **334**, 1091–1097 (2011).

82. F. Delsuc, *et al.*, A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biol* **16**, 39 (2018).

83. K. M. Cohen, S. C. Finney, P. L. Gibbard, J.-X. Fan, The ICS International Chronostratigraphic Chart. *Episodes Journal of International Geoscience* **36**, 199–204 (2013).

Figure Legends

Figure 1: Cluster Analysis and Phylogeny of Ligand groups. A) Similarity-based clustering, using CLANS, of canonical chemokines and related molecules with sequence similarity. Canonical chemokines are an independent group from other related molecules (TAFA, CYTL and CXCL17). Canonical chemokines are composed of two large groups (CC-type and CXC-type) within which some

divergent subgroups are highlighted. The clustering and connections shown are at the p-value threshold of 1E-6. Other p-values tested are shown in Supplementary Figure S1. Candidate invertebrate sequences are shown as crosses and further information regarding them can be found in Supplementary Results. **B)** Maximum-Likelihood un-rooted phylogenetic tree of canonical chemokines. CC-type and CXC-type are split into two separate clades. Supports for key nodes are indicated in boxes with Transferable Bootstrap Expectation (TBE) represented by triangles and the Ultrafast Bootstraps (UFB) as circles. A traffic light colour code is used to indicate the level of support: high (green); intermediate (yellow) and low (red). **C)** Similarity-based clustering, using CLANS, of the CKLF super family (CKLFSF). Two major clusters are formed: the smaller "CKLF Group I" and the heterogenous "CKLF group II" that also includes some invertebrate sequences (shown as crosses). Subclades, including the known members of the CKLF super family, are highlighted. The clustering and connections shown are at the p-value threshold of 1E-15, as this is the threshold at which the two major clusters connect. Other p-values tested are shown in Supplementary Figure S2. **D)** Maximum-Likelihood un-rooted phylogenetic tree of the CKLF super family (CKLFSF). The CKLF group I is monophyletic, while the CKLF group II is not. Supports for key nodes are indicated in boxes with Transferable Bootstrap Expectation (TBE) represented by triangles and the Ultrafast Bootstraps (UFB) as circles. A traffic light colour code is used to indicate the level of support: high (green); intermediate (yellow) and low (red).

Figure 2: Distribution and duplication patterns of ligand groups. A) Presence of all ligand groups are mapped onto a species tree. Gene trees and duplication events are based on the gene tree to species tree reconciliation analyses. The nomenclature for canonical chemokines is primarily based on known chemokines of human (or mouse). Where human and mouse chemokines do not correspond, the default name refers to the human gene and the mouse (*Mus musculus*) one is indicated with "Mm". Chemokines that have been classically described as having either homeostatic or inflammatory function are indicated with a circle or a star respectively. The classification used

here was based on Zlotnik and Yoshie 2012 (30) with the inflammatory type also including chemokines they described as plasma/platelet types. Overall, canonical chemokines originated in vertebrates and expanded a first time in jawed vertebrates and a second time in mammals. Homeostatic chemokines (e.g., CXCL12) are generally more ancient than inflammatory ones. CXCL17 and CYTL are mammal and jawed vertebrate specific respectively. TAFA originated in the common ancestor of vertebrates and urochordates, while the CKLF super family is present in invertebrates although key duplications occurred at the base of vertebrates. **B)** Number of complements for each ligand group at key species nodes are mapped onto the species tree. The number of complements in each group reflects the pattern of duplications. The major increase occurred at the level of jawed vertebrates with canonical chemokines undergoing a second significant increase within placentals.

Figure 3: Phylogeny of Receptor groups.

An unrooted maximum likelihood phylogeny of chemokine receptors. The tree shown is the transfer bootstrap expectation (TBE) tree including just the chordate specific clade from the ultrafast bootstrap tree (UFB). Node supports from both TBE (triangle) and UFB (circle) shown for equivalent key nodes in boxes with arrows to indicate node. A traffic light colour code is used to indicate the level of support: high (green); intermediate (yellow) and low (red). Key clades highlighted: yellow = chemokine like plus group (CMLplus); blue = intermediate group; green = atypical 3 and GPR182 (ACKR3/GPR182); purple = canonical chemokines (Canonical CKR); and pink = relaxin receptors (RL3R). Branches scaled by amino acid substitutions per site.

Figure 4: Distribution and duplication patterns of receptor groups.

A) Presence of all receptor groups are mapped onto a species tree. Gene trees and duplication events are based on the gene tree to species tree reconciliation analyses. The nomenclature for genes is primarily based on human chemokines. The canonical chemokines had 5 paralogs present in the vertebrate common ancestor. These undergo a heterogeneous pattern of duplication

throughout vertebrates with different paralogs duplicating different number of times and in different groups of species. Chemokines that have been classically described as having either homeostatic or inflammatory function are indicated with a circle or a star respectively. The classification used here was based on Zlotnik and Yoshie 2012 (30). **B)** Number of complements for each receptor group at key species nodes are mapped onto the species tree. The number of complements in each group reflects the pattern of duplications. The chemokine groups diverged in the vertebrate stem group. The major expansion occurred at the level of jawed vertebrates with canonical chemokine receptors, the chemokine-like receptor plus group and intermediate groups increasing in copy number. Canonical chemokine underwent another small subsequent increase within placentals.

Figure 5: Summary of the evolution of ligand and receptors.

A summary diagram of the evolution of the different chemokine system components. A simplified phylogenetic tree of species is shown, calibrated to time according to Erwin et al. 2011 (81), Delsuc et al. 2018 (82)and Cohen et al. 2013 (83). Circles represent ligand groups and 7 transmembrane domain structure icon represents GPCR groups. Icons are colour coded by group and placed adjacent to the branch in the species tree where they first appear. X2 and X5 shown to indicate the number of paralogs present for CXCL ligand group and the canonical CKR groups respectively, on the branch where they first appear. Question mark refers to the uncertainty regarding the origin of the CKLF group I in jawed vertebrates or deuterostome stem group (see Figure 2). Geological column is shown along the bottom, in accordance with the ICS International Chronostratigraphic Chart.

Supplementary Tables:

Table S1. Summary of canonical and non-canonical chemokine system components.

Table S2. List of the species used in this work and their respective BUSCO values and accession number.

Table S3. Substitution models used in the phylogenetic analyses.

Supplementary Figure captions:

Figure S1. CLANS clustering of chemokines and related molecules sequences. Initial identification and annotation of clusters was performed at the strict p-value of 1E-35 **(A)**. Subsequent loosening of the p-value clarified the relationships across clusters and defined bigger groups. At p-value 1E-15 **(B)** two major canonical chemokine groups are well-defined: the CCL group, which includes also XCL and X3CL; and the CXCL group. At this level of stringency, only few canonical chemokines remain isolated: CCL27/28; CXCL12; CXCL14; CXCL16; as well as CXCL17. TAFA and CYTL are also isolated. At p-value 1E-10 **(C)** the two major chemokine groups connect to each other. CCL27/28 is connected to the CCL group and CXCL12 and CXCL14 are connected to the CXCL group, while CXCL16 and CXCL17 are still isolated. At p-value 1E-6 **(D)** all chemokine groups are connected in one big cluster, except for CXCL17. TAFA and CYTL are also still isolated. Crosses indicate the few invertebrate sequences that were collected from the BLAST search, more information in Supplementary Results.

Figure S2. CLANS clustering of CKLF Super Family sequences. Initial identification and annotation of clusters was performed at the strict p-value of 1E-60 **(A)**. Subsequent loosening of the p-value clarified the relationships across clusters and defined bigger groups. At 1E-20 **(B)** two major clusters have formed. One, that we called CKLF group I, includes CKLF, CMTM1,2,3,5, and PLP2. The other, that we called CKLF group II, includes CMTM4/6,7,8, and other groups. At 1E-16 **(C)** more sequences have joined the two major groups that are still separate. At 1E-15 **(D)** the two major groups connect,

as well as few extra sequences, see Supplementary Results for extra details. Crosses indicate invertebrate sequences.

Figure S3. Alignment of candidate brachiopod CCL24 sequence with mammalian CCL24s. Our BLAST searches picked up a sequence from the brachiopod *Lingula unguis* that when re-blasted versus SwissProt returned a CCL24 as hit. Alignment of the brachiopod sequence with mammalian CCL24 sequences reveals a poor overall conservation, with the brachiopod sequence also being significantly longer than any of the other sequences. Further details about this sequence can be found in supplementary file S3 and in Supplementary Results.

Figure S4. Alignment of candidate cnidarian CCL3 sequence with mammalian CCL3s. Our BLAST searches picked up a sequence from the cnidarian *Clytia hemisphaerica* that when re-blasted versus SwissProt returned a CCL3 as hit. Alignment of the cnidarian sequence with mammalian CCL3 sequences reveals a poor overall conservation, with the cnidarian sequence being extremely longer than any of the other sequences. Further details about this sequence can be found in supplementary file S3 and in Supplementary Results.

Figure S5. Alignment of candidate echinoderm CXCL10 sequence with mammalian CXCL10s. Our BLAST searches picked up a sequence from the echinoderm *Acanthaster planci* that when re-blasted versus SwissProt returned a CXCL10 as hit. Alignment of the echinoderm sequence with mammalian CXCL10 sequences reveals a poor overall conservation, with the brachiopod sequence also being significantly longer than any of the other sequences. Further details about this sequence can be found in supplementary file S3 and in Supplementary Results.

Figure S6. Alignment of 4 candidate urochordate TAFA sequences with vertebrate TAFAs. Our BLAST searches picked up 4 sequences from the urochordate *Ciona intestinalis*, that connected with

the TAFA cluster in the CLANS analysis. One of these sequences when blasted versus SwissProt returned a TAFA as hit. This sequence was also annotated as TAFA by InterProScan. Alignment of the urochordate sequences with vertebrate TAFA sequences reveals that only the one annotated as TAFA aligns well. While the other 3 align poorly and are also significantly longer than any of the other sequences. Further details about these sequences can be found in supplementary file S3 and in Supplementary Results.

Figure S7. Alignment of best candidate urochordate TAFA sequence with vertebrate TAFAs. Of the 4 urochordate candidate TAFA sequences, only one was annotated as TAFA with both SwissProt and InterProScan annotation and appeared to align well with other TAFAs with a preliminary alignment with all urochordate sequences (Figure S6). Here we removed the other 3 urochordate sequences and aligned only the best candidate with the vertebrate TAFAs. The sequence conservation is even more apparent with this alignment. Importantly, 8 of the 10 typical cysteine residues of TAFA1-4 are conserved, and the two missing cysteines are the same ones missing in TAFA5. Further discussion can be found in Supplementary Results.

Figure S8. Unrooted phylogenetic tree of canonical chemokines with TBE supports. Phylogenetic tree under the model GTR20+F+R4. Nodal support is calculated from 100 non-parametric bootstrap repeats with transferable bootstrap expectation. CCL clade is in orange, CXCL clade in blue.

Figure S9. Unrooted phylogenetic tree of canonical chemokines with UFB supports. Phylogenetic tree under the model GTR20+F+R4. Nodal support is calculated from 1000 ultrafast bootstrap repeats. CCL clade is in orange, CXCL clade in blue.

Figure S10. Unrooted phylogenetic tree of TAFA with TBE supports. Phylogenetic tree under the model JTT+R5. Nodal support is calculated from 100 non-parametric bootstrap repeats with transferable bootstrap expectation.

Figure S11. Unrooted phylogenetic tree of TAFA with UFB supports. Phylogenetic tree under the model JTT+R5. Nodal support is calculated from 1000 ultrafast bootstrap repeats.

Figure S12. Unrooted phylogenetic tree of CYTL with TBE supports. Phylogenetic tree under the model JTT+I+G4. Nodal support is calculated from 100 non-parametric bootstrap repeats with transferable bootstrap expectation.

Figure S13. Unrooted phylogenetic tree of CYTL with UFB supports. Phylogenetic tree under the model JTT+I+G4. Nodal support is calculated from 1000 ultrafast bootstrap repeats.

Figure S14. Unrooted phylogenetic tree of CXCL17 with TBE supports. Phylogenetic tree under the model JTT. Nodal support is calculated from 100 non-parametric bootstrap repeats with transferable bootstrap expectation.

Figure S15. Unrooted phylogenetic tree of CXCL17 with UFB supports. Phylogenetic tree under the model JTT. Nodal support is calculated from 1000 ultrafast bootstrap repeats.

Figure S16. Unrooted phylogenetic tree of CKLFSF with TBE supports. Phylogenetic tree under the model GTR20+F+R7. Nodal support is calculated from 100 non-parametric bootstrap repeats with transferable bootstrap expectation. Red clade = CMTM4/6; blue clade = CKLF I group; green clade = CMTM7; turquois clade = MAL/MALL/MAL2.

Figure S17. Unrooted phylogenetic tree of CKLFSF with UFB supports. Phylogenetic tree under the model GTR20+F+R7. Nodal support is calculated from 1000 ultrafast bootstrap repeats. Red clade = CMTM4/6; blue clade = CKLF I group; green clade = CMTM7; turquois clade = MAL/MALL/MAL2.

Figure S18. Rooted species tree reconciled gene tree for canonical chemokines. The canonical chemokines gene tree was reconciled with the species tree using GeneRax. "S" or "D" at the node indicates a speciation or duplication event respectively. CCL clade is in orange, CXCL clade in blue.

Figure S19. Rooted species tree reconciled gene tree for TAFA. The TAFA gene tree was reconciled with the species tree using GeneRax. "S" or "D" at the node indicates a speciation or duplication event respectively.

Figure S20. Rooted species tree reconciled gene tree for CYTL. The CYTL gene tree was reconciled with the species tree using GeneRax. "S" or "D" at the node indicates a speciation or duplication event respectively.

Figure S21. Rooted species tree reconciled gene tree for CXCL17. The CXCL17 gene tree was reconciled with the species tree using GeneRax. "S" or "D" at the node indicates a speciation or duplication event respectively.

Figure S22. Rooted species tree reconciled gene tree for CKLFSF. The CKLFSF gene tree was reconciled with the species tree using GeneRax. "S" or "D" at the node indicates a speciation or duplication event respectively. Red clade = CMTM4/6; blue clade = CKLF I group; green clade = CMTM7; turquois clade = MAL/MALL/MAL2.

Figure S23. CLANS clustering of receptors and related molecules sequences. A CLANS clustering layout where shapes indicate sequences and lines are connections indicating similarity between sequences at or surpassing the p-value similarity threshold. Sequences are positioned in clusters based on similarity. Initial identification and annotation of clusters was performed using the inbuilt convex clustering at the p-value of 1E-100. **(A)** clustering was loosened till the canonical receptor annotated groups formed a cluster at 1E-65. **(B)** loosening of the p-value to 1E-60 identified relationships between clusters of interest and identified the intermediate group as connecting to both canonical and chemokine-like plus groups. All sequences connected to groups of interest are vertebrate sequences. **(C)** further loosening to p-value 1E-50 connects the vertebrate sequences of interest to a large cluster of sequences which contains vertebrate and invertebrate sequences which are annotated as opioid and somatostatin receptors and other GPCRs. Crosses indicate invertebrate sequences and Y-shape indicates the reference viral sequences included. Shapes are colour coded by group of interest: purple = canonical chemokine receptors; yellow = chemokine-like plus; green = atypical receptor 3/GPR182; blue = intermediate group; pink = relaxin receptors.

Figure S24. Unrooted phylogenetic tree of receptors with TBE supports. Phylogenetic tree of receptor sequences of interest and putative outgroups under the model GTR20+F+G4. Sequences used are a subset of sequences extracted from CLANs, specifically they are those in the chordate specific clade in the ultrafast bootstrap tree. Nodal support is calculated from 100 non-parametric bootstrap repeats with transferable bootstrap expectation. Branches colour coded by group of interest: purple = canonical chemokine receptors; yellow = chemokine-like plus; green = atypical receptor 3/GPR182; blue = intermediate group; pink = relaxin receptors.

Figure S25. Unrooted phylogenetic tree of receptors with UFB supports. Phylogenetic tree of all receptor sequences of interest and putative outgroups extracted from clans under the model GTR20+F+G4. Nodal support is calculated from 1000 ultrafast bootstrap repeats. Branches colour

coded by group of interest: purple = canonical chemokine receptors; yellow = chemokine-like plus; green = atypical receptor 3/GPR182; blue = intermediate group; pink = relaxin receptors.

Figure S26. Rooted species tree reconciled gene tree for receptors. The ultrafast bootstrap receptor tree was modified to extract the subtree of the chordate specific clade. This gene tree was reconciled with the species tree using GeneRax. “S” or “D” at the node indicates a speciation or duplication event respectively. Branches colour coded by group of interest: purple = canonical chemokine receptors; yellow = chemokine-like plus; green = atypical receptor 3/GPR182; blue = intermediate group; pink = relaxin receptors.

Fig. 1

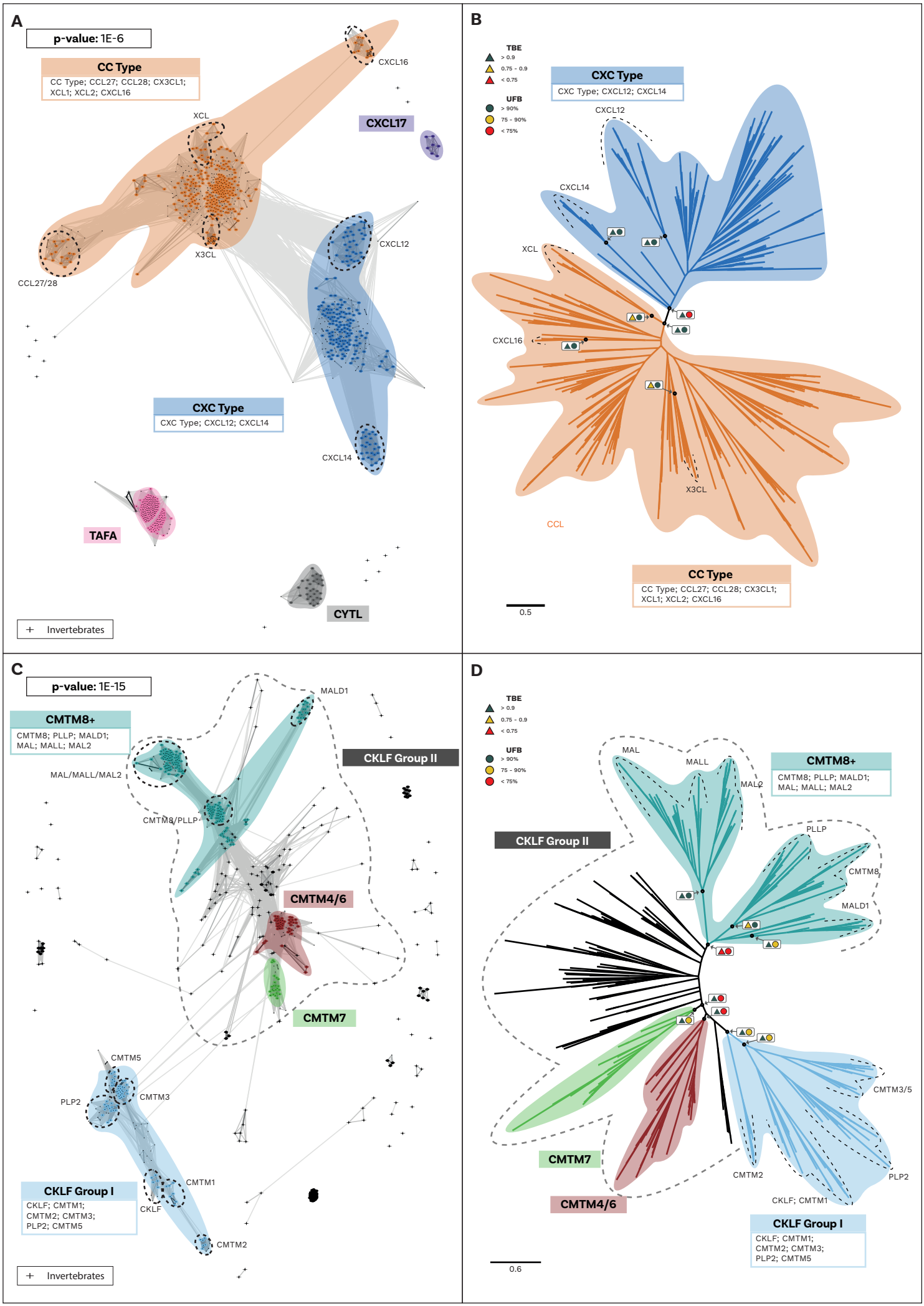

A
p-value: 1E-6
CC Type
CC Type; CCL27; CCL28; CX3CL1; XCL1; XCL2; CXCL16
CXCL16
XCL
CXCL17
CXCL12
X3CL
CCL27/28
CXC Type
CXC Type; CXCL12; CXCL14
CXCL14
TAFA
CYTL
+ Invertebrates
B
TBE
> 0.9
0.75 - 0.9
< 0.75
UFB
> 90%
75 - 90%
< 75%
CXC Type
CXC Type; CXCL12; CXCL14
CXCL12
CXCL14
XCL
CXCL16
X3CL
CCL
CC Type
CC Type; CCL27; CCL28; CX3CL1; XCL1; XCL2; CXCL16
0.5
C
p-value: 1E-15
MALD1
CMTM8+
CMTM8; PLLP; MALD1; MAL; MALL; MAL2
CKLF Group II
MAL/MALL/MAL2
CMTM8/PLLP
CMTM4/6
CMTM7
CMTM5
PLP2
CMTM3
CMTM1
CKLF
CMTM2
CKLF Group I
CKLF; CMTM1; CMTM2; CMTM3; PLP2; CMTM5
+ Invertebrates
D
TBE
> 0.9
0.75 - 0.9
< 0.75
UFB
> 90%
75 - 90%
< 75%
MAL
MALL
MAL2
CMTM8+
CMTM8; PLLP; MALD1; MAL; MALL; MAL2
CKLF Group II
PLLP
CMTM8
MALD1
CMTM3/5
CMTM7
CMTM2
PLP2
CMTM4/6
CKLF; CMTM1
CKLF Group I
CKLF; CMTM1; CMTM2; CMTM3; PLP2; CMTM5
0.6

Fig. 2

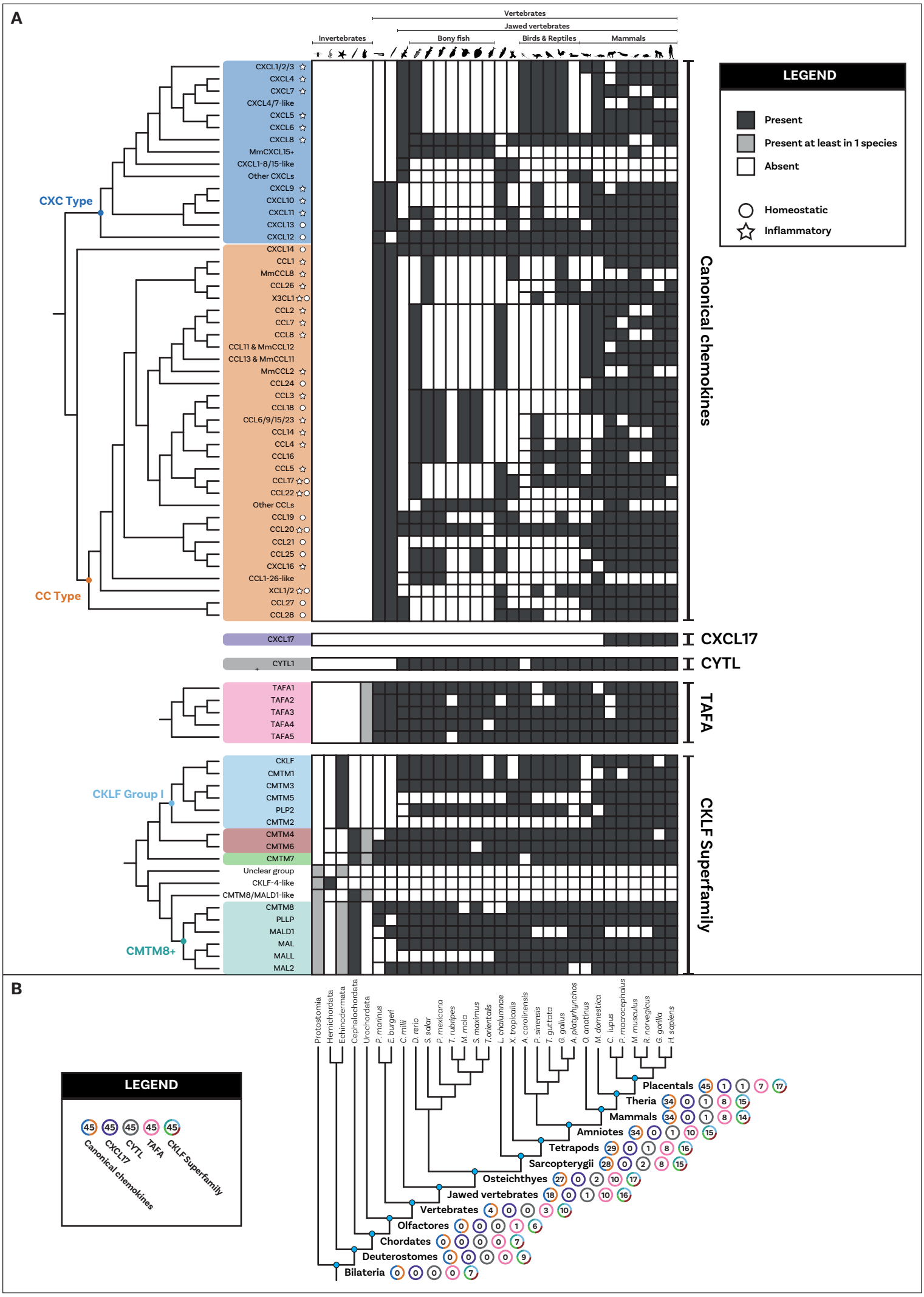

A
Vertebrates
Jawed vertebrates
Invertebrates
Bony fish
Birds & Reptiles
Mammals
CXCL1/2/3
CXCL4
CXCL7
CXCL4/7-like
CXCL5
CXCL6
CXCL8
MmCXCL15+
CXCL1-8/15-like
Other CXCLs
CXCL9
CXCL10
CXCL11
CXCL13
CXCL12
CXCL14
CCL1
MmCCL8
CCL26
X3CL1
CCL2
CCL7
CCL8
CCL11 & MmCCL12
CCL13 & MmCCL11
MmCCL2
CCL24
CCL3
CCL18
CCL6/9/15/23
CCL14
CCL4
CCL16
CCL5
CCL17
CCL22
Other CCLs
CCL19
CCL20
CCL21
CCL25
CXCL16
CCL1-26-like
XCL1/2
CCL27
CCL28
CXC Type
CC Type
Canonical chemokines
CXCL17
CYTL1
CXCL17
CYTL
TAFA1
TAFA2
TAFA3
TAFA4
TAFA5
TAFA
CKLF
CMTM1
CMTM3
CMTM5
PLP2
CMTM2
CMTM4
CMTM6
CMTM7
Unclear group
CKLF-4-like
CMTM8/MALD1-like
CMTM8
PLLP
MALD1
MAL
MALL
MAL2
CKLF Group I
CMTM8+
CKLF Superfamily
LEGEND
Present
Present at least in 1 species
Absent
Homeostatic
Inflammatory
B
Protostomia
Hemichordata
Echinodermata
Cephalochordata
Urochordata
P. marinus
E. burgeri
C. milii
D. rerio
S. salar
P. mexicana
T. rubripes
M. mola
S. maximus
T. orientalis
L. chalumnae
X. tropicalis
A. carolinensis
P. sinensis
T. guttata
G. gallus
A. platyrhynchos
O. anatinus
M. domestica
C. lupus
P. macrocephalus
M. musculus
R. norvegicus
G. gorilla
H. sapiens
Placentals 45 1 1 7 17
Theria 34 0 1 8 15
Mammals 34 0 1 8 14
Amniotes 34 0 1 10 15
Tetrapods 29 0 1 8 16
Sarcopterygii 28 0 2 8 15
Osteichthyes 27 0 2 10 17
Jawed vertebrates 18 0 1 10 16
Vertebrates 4 0 0 3 10
Olfactores 0 0 0 1 6
Chordates 0 0 0 0 7
Deuterostomes 0 0 0 0 9
Bilateria 0 0 0 0 7
LEGEND
45 Canonical chemokines
45 CXCL17
45 CYTL
45 TAFA
45 CKLF Superfamily

Fig. 3

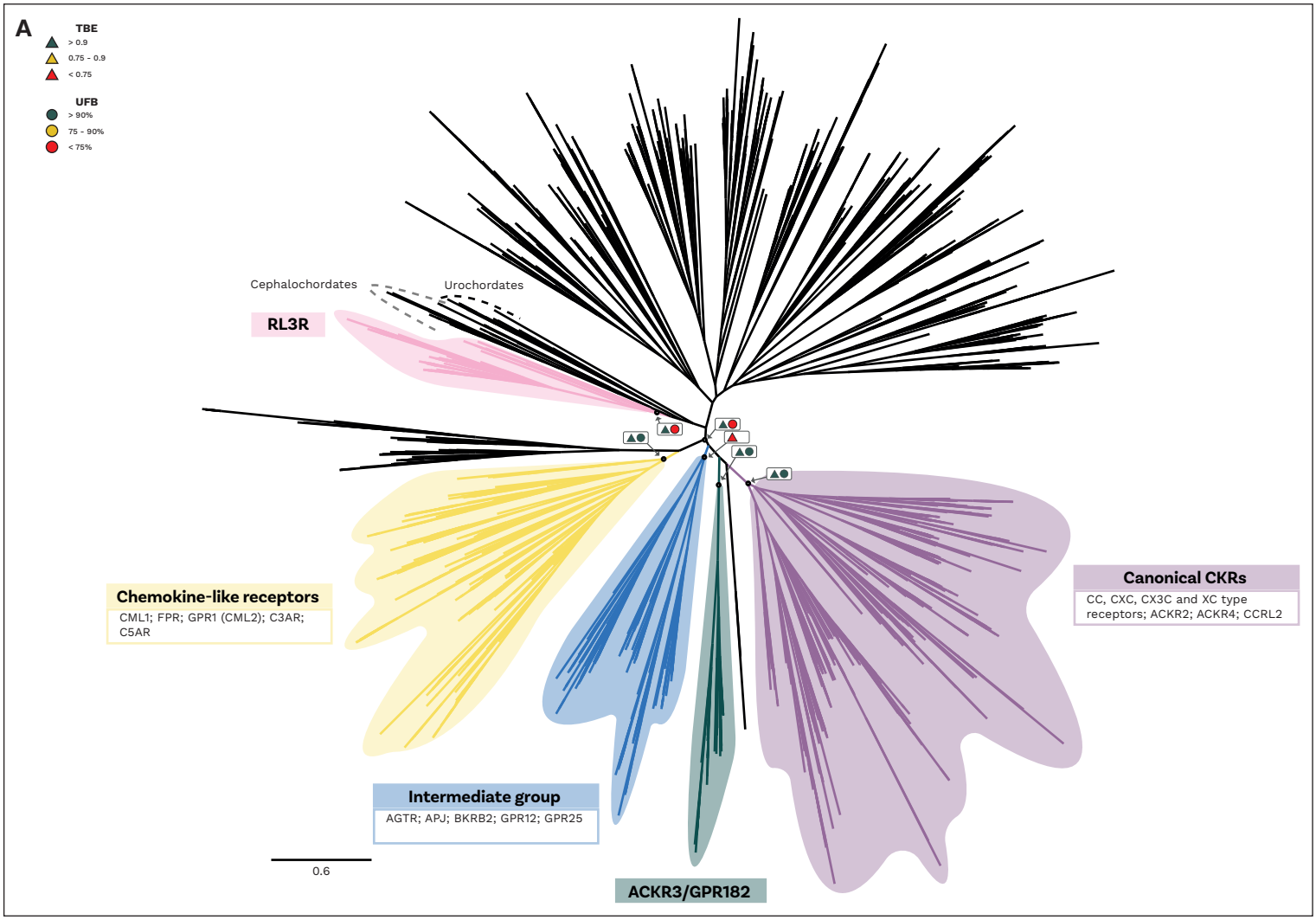

A
TBE
> 0.9
0.75 - 0.9
< 0.75
UFB
> 90%
75 - 90%
< 75%
Cephalochordates
Urochordates
RL3R
Chemokine-like receptors
CML1; FPR; GPR1 (CML2); C3AR; C5AR
Intermediate group
AGTR; APJ; BKRB2; GPR12; GPR25
ACKR3/GPR182
Canonical CKRs
CC, CXC, CX3C and XC type receptors; ACKR2; ACKR4; CCRL2
0.6

Fig. 4

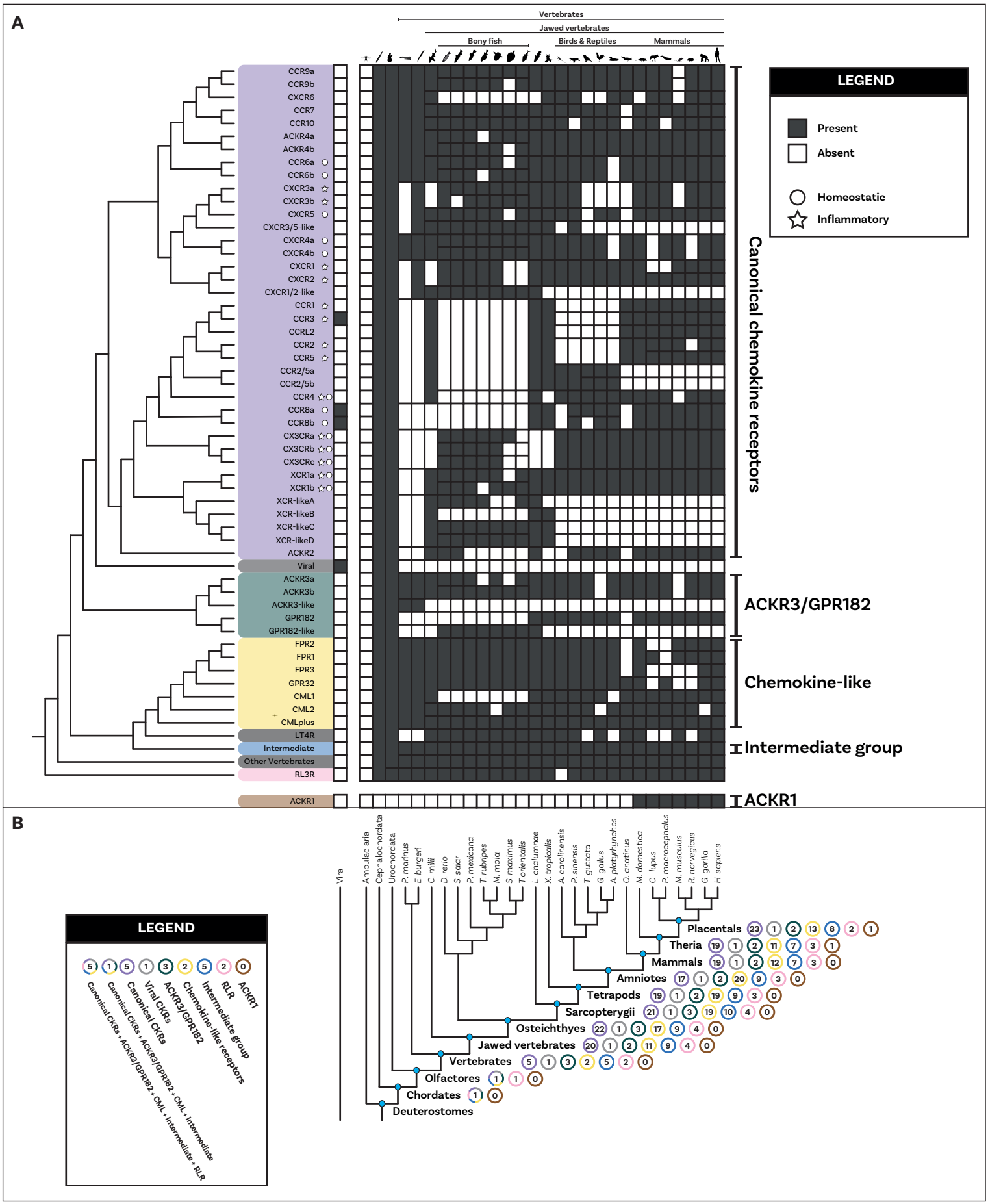
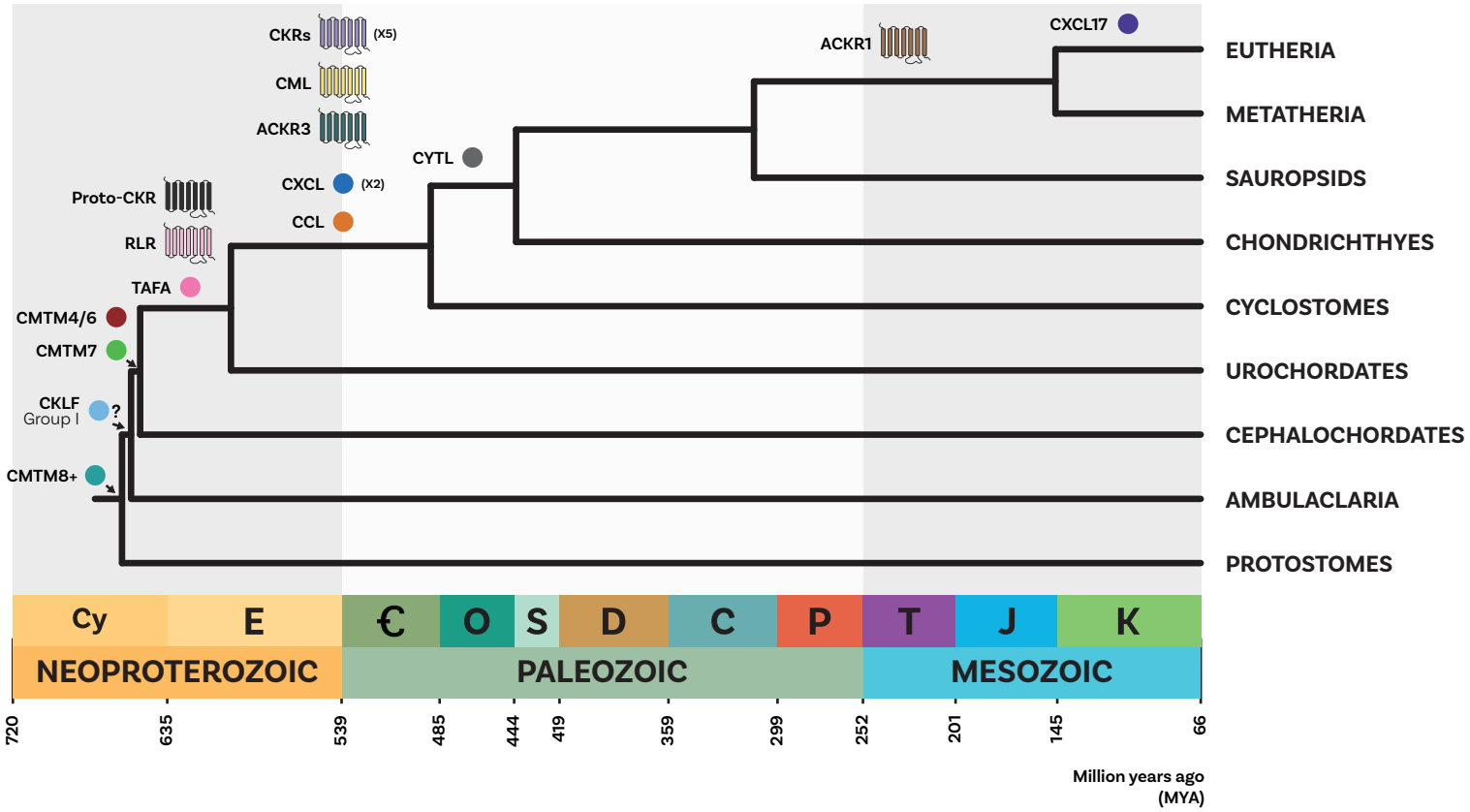

A
Vertebrates
Jawed vertebrates
Bony fish
Birds & Reptiles
Mammals
CCR9a
CCR9b
CXCR6
CCR7
CCR10
ACKR4a
ACKR4b
CCR6a
CCR6b
CXCR3a
CXCR3b
CXCR5
CXCR3/5-like
CXCR4a
CXCR4b
CXCR1
CXCR2
CXCR1/2-like
CCR1
CCR3
CCRL2
CCR2
CCR5
CCR2/5a
CCR2/5b
CCR4
CCR8a
CCR8b
CX3CRa
CX3CRb
CX3CRc
XCR1a
XCR1b
XCR-likeA
XCR-likeB
XCR-likeC
XCR-likeD
ACKR2
Viral
ACKR3a
ACKR3b
ACKR3-like
GPR182
GPR182-like
FPR2
FPR1
FPR3
GPR32
CML1
CML2
CMLplus
LT4R
Intermediate
Other Vertebrates
RL3R
ACKR1
LEGEND
Present
Absent
Homeostatic
Inflammatory
Canonical chemokine receptors
ACKR3/GPR182
Chemokine-like
Intermediate group
ACKR1
B
Viral
Ambulacraria
Cephalochordata
Urochordata
P. marinus
E. burgeri
C. milii
D. rerio
S. salar
P. mexicana
T. rubripes
M. mola
S. maximus
T. orientalis
L. chalumnae
X. tropicalis
A. carolinensis
P. sinensis
T. guttata
G. gallus
A. platyrhynchos
O. anatinus
M. domestica
C. lupus
P. macrocephalus
M. musculus
R. norvegicus
G. gorilla
H. sapens
Placentals 23 1 2 13 8 2 1
Theria 19 1 2 11 7 3 1
Mammals 19 1 2 12 7 3 0
Amniotes 17 1 2 20 9 3 0
Tetrapods 19 1 2 19 9 3 0
Sarcopterygii 21 1 3 19 10 4 0
Osteichthyes 22 1 3 17 9 4 0
Jawed vertebrates 20 1 2 11 9 4 0
Vertebrates 5 1 3 2 5 2 0
Olfactores 1 1 0
Chordates 1 0
Deuterostomes
LEGEND
5 1 5 1 3 2 5 2 0
Canonical CKRs + ACKR3/GPR182 + CML + Intermediate + RLR
Canonical CKRs + ACKR3/GPR182 + CML + Intermediate
Canonical CKRs
Viral CKRs
ACKR3/GPR182
Chemokine-like receptors
Intermediate group
RLR
ACKR1

Fig. 5


CKRs
(X5)
CML
ACKR3
CXCL
(X2)
CCL
CYTL
Proto-CKR
RLR
TAFA
CMTM4/6
CMTM7
CKLF
Group I
?
CMTM8+
ACKR1
CXCL17
EUTHERIA
METATHERIA
SAUROPSIDS
CHONDRICHTHYES
CYCLOSTOMES
UROCHORDATES
CEPHALOCHORDATES
AMBULACLARIA
PROTOSTOMES
Cy
E
Є
O
S
D
C
P
T
J
K
NEOPROTEROZOIC
PALEOZOIC
MESOZOIC
720
635
539
485
444
419
359
299
252
201
145
66
Million years ago
(MYA)