Chapter 5

The Origin, Evolution and Molecular Diversity of the Chemokine System

Preface

The work presented in this chapter has been published as a pre-print (available on bioRxiv[1]) and has undergone a first round of reviewing at the independent reviewing platform Review Commons[2]. The version of the manuscript presented here has been reformatted to match the style of the rest of this thesis.

This work was done in collaboration with other members of the Feuda Group that are also co-authors in the pre-print. Specifically, I was the lead in the evolution of chemokine ligands, Matthew Goulty was the lead in the evolution of chemokine receptors, Clifton Lewis contributed to mapping the evolution of all chemokine components to a calibrated species tree and was also heavily involved in structuring the figures, Flaviano Giorgini provided useful comments throughout the manuscript preparation and Roberto Feuda proposed the original idea of the project and supervised the work.

1. https://www.biorxiv.org/content/10.1101/2023.05.17.541135v2.full
2. https://www.reviewcommons.org/about/

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. [#] Corresponding authors. [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. Several non-canonical components interact with the system. These include a variety of molecules that usually share some degree of sequence similarity with canonical components and, in some cases, are known to bind to canonical components and/or to modulate cell migration (1, 2). 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 (3–5). The system has two components, a ligand, usually a small cytokine called a chemokine, and a receptor. It typically operates through chemoattraction, wherein one cell type produces and secretes chemokines, creating a chemical gradient as these molecules disperse. Cells equipped with the corresponding chemokine receptors on their membranes can recognize and bind to specific chemokines, promoting their migration along the gradient (4). This mechanism allows cells to reach target locations, such as infection sites during inflammation or tissues important for homeostatic functions, e.g., leukocyte maturation and trafficking (3, 6). Chemokines involved in the latter homeostatic functions are usually constitutively expressed, while those involved in inflammatory responses have an inducible expression (7). Chemokine 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 (8). 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 (9). In addition to canonical components, other molecules have been discovered to function similarly to chemokine ligands (1) or receptors (2) (see Table 5.1). These include: the chemokine-like factor (CKLF) that binds to chemokine receptor CCR4 (10, 11) and drives cell migration *in vivo* (12); TAFA chemokines, expressed mainly in the nervous system, which share structural similarities to canonical chemokines (13, 14) and bind GPCRs related to chemokine receptors, e.g. formyl peptide receptors (15, 16) and GPR1 (17); Cytokine-like 1 (CYTL1) that binds CCR2 (18) and has been suggested to be related to CC ligands based on the presence of a IL8-like chemokine fold (19). There are also non-canonical chemokine receptors, such as: the chemokine-like receptor (CML1, or also CMKLR1) ((20); atypical chemokine receptors (ACKRs) (21); and viral chemokine receptors (22–25). Unlike other chemokine receptors, atypical receptors cannot initiate classical chemokine signaling upon ligand binding (21, 26). 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) (27, 28). Additionally, several proteins of viral origins, such as US28

from human cytomegalovirus, have chemokine-receptor/binding activity (22, 23). These viral proteins can bind a wide array of chemokine ligands (23).

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 (12, 13, 19, 29). 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 (9, 30, 31). 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 (7, 32)). 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 (33).

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.

Table 5.1. Summary table of all the canonical and non-canonical chemokine components analyzed in this study.

	Names	Abbreviations	*H. sapiens* Orthologs	Functions	References
Ligand Groups	**Canonical Chemokines**	CCL, CXCL, XCL, CX3CL	CCL1-3, 3L1,L3, 4, 4L1-L2, 5, 7, 8, 11, 13, 14-28; CXCL1-4, 4L1, 5-14, 16,17; XCL1,2; CX3CL1	- Chemokine receptor binding and signalling - Chemoattraction of leukocytes - Homeostasis of leukocytes	(2, 4, 7)
	CKLF-Like MARVEL Transmembrane Domain-Containing Proteins (Chemokine-Like Factor Super Family)	CKLF, CMTM	CKLF1; CMTM1-8 (CKLF, CKLFSF1-8)	- CKLF1 (CKLF) binds to chemokine receptor CCR4 - CKLF1 (CKLF): chemotactic activity for lymphocytes, macrophages, and neutrophils - Other CMTMs: variably expressed in immune system; putative roles in immunity, programmed cell death, regulation of anti-tumour immunity etc.	(1, 10–12, 34–42)
	Cytokine-Like Protein 1 (Protein C17 or C4orf4)	CYTL	CYTL1	- Chemokine receptor binding (CCR2) and signalling - Chemoattraction monocytes/macrophages - Chemotactic activity in neutrophils	(1, 18, 43, 44)
	TAFA Chemokines (Family with sequence similarity 19 (chemokine (C-C motif)-like) member A)	TAFA	TAFA1-5 (FAM19A1-5)	- Formyl-peptide receptor binding and signalling (TAFA4 and 5) - Putative binding to other GPCRs: GPR1 (TAFA1); S1PR2 (TAFA5) - Expressed in central and peripheral nervous system - Implicated in vast diversity of physiological processes	(1, 13–17, 45–47)
Receptor Groups	**Canonical Chemokine Receptors**	CCR, CXCR, XCR, CX3CR	CCR1-10; CXCR1-6; XCR1; CX3CR1	- Chemokine binding and signalling - Chemotaxis of leukocytes - Homeostasis of leukocytes	(2, 4, 7)
	Atypical Chemokine Receptors	ACKR	ACKR1-4 (DARC; D6; CXCR7; CCRL1)	- Chemokine binding, but no signalling - Resolution of inflammatory response	(21, 27, 28)
	Chemokine Receptor-Like (Chemokine C-C motif receptor-like2)	CCRL	CCRL2 (ACKR5)	- Binds CCL5 and CCL19, but no signalling - Binds chemerin and presents it to CMKLR1	(20, 48)
	Chemokine-Like Receptor 1	CML	CML1 (CMKLR1; ChemR23)	- Binds chemerin inducing migration of macrophages and dendritic cells - Binds also other anti-inflammatory molecules (e.g., Resolvin E1 (RvE1))	(20)
	Formyl-peptide Receptors	FPR	FPR 1-3	- TAFA chemokine binding - Chemoattraction, modulation of inflammation	(15, 16, 49)
	Putative Chemokine Receptors	ACKR6, CXCR8	PTITMP3, CXCR8 (GPR35)	- ACKR6/PTITMP3: Binds CCL18 (NB: It is not a GPCR) - CXCR8/GPR35: binds CXCL17	(48, 50)

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 (Table 5.1). The presence of a four transmembrane MARVEL domain in the latter proteins (12, 34, 35) distinguishes them from canonical chemokines, the CYTL and the TAFAs. Therefore, we separated these two groups for further analysis. Using BLASTP or PSI-BLAST (51–53) (see Materials and Methods for more details) against 64 species from 19 animal phyla (Table S1), we identified 891 putative homologs for chemokines, TAFA and CYTL and 602 putative homologs of the CKLF Super Family.

We utilized CLANS (54, 55), 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 (54). 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 5.1A). Whereas, CXCL17, TAFA and CYTL remain separate from canonical chemokines and from each other even at the loosest p-values tested (Figure 5.1A). The distinction between CXCL17 and all other canonical chemokines is consistent with our receptor results showing that the potential receptor for CXCL17, GPR35 (50), is also not within the canonical chemokine receptor group (see below). Although it is important to note that recent studies fail to demonstrate CXCL17 activity at GPR35 (56, 57). Within the CKLFSF, two large clusters were identified, named CKLF I and CKLF II, although these ultimately connect to form one large superfamily (Figure 5.1B). 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 5.1A and 5.1B).

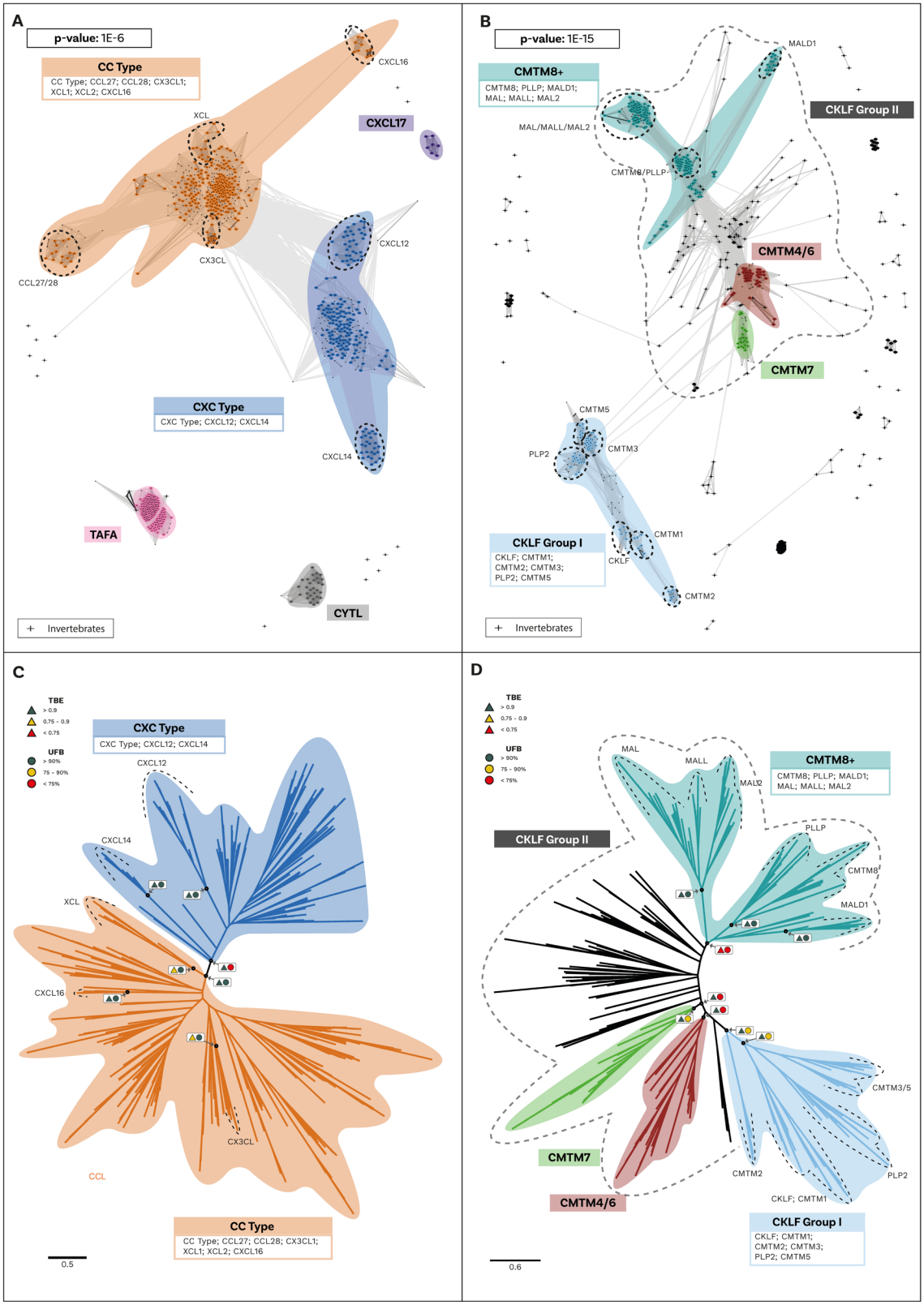

Figure 5.1. Cluster Analysis and Phylogeny of Ligand groups. (A) Similarity-based clustering, using CLANS, of canonical chemokines and related molecules with sequence similarity. Canonical chemokines

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)** 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. **(C)** 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). **(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).

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 5.1C, D and Figures S8-17) (see methods for details). To evaluate the nodal support, in addition to the UltraFast bootstrap (UFB) (58, 59), we used Transfer Bootstrap Expectation (TBE), a method that has been developed for single-gene phylogeny (60). To evaluate ortholog/paralog relationships and overall dynamics of the ligand complement, we used GeneRax (61). This new method uses maximum likelihood to reconcile the gene tree with the species tree (61). In brief, given a gene and species tree, GeneRax uses a maximum likelihood approach to optimize the duplication and loss events (61, 62).

Our analysis initially identified a few invertebrate putative chemokine ligands (Figure 5.1A), however, these sequences lacked protein signatures associated with the canonical ligands (Figures S3-5 and File S3), and they were therefore excluded from further analysis (see Supplementary Results for further information). The phylogenetic tree for the canonical ligands identifies two major groups, the CC-type, which also includes the XC- and X3C-types, and the CXC-type (TBE = 0.95, UFB = 92%) (Figure 5.1C and Figures S8,9), confirming the previous finding obtained using synteny data (63, 64). Next, to clarify the distribution of canonical chemokines, we first reconciled their gene tree with the species tree and then used the reconciled tree to trace the

presence/absence of each chemokine group throughout all the species (Figure 5.2A and Figure S18). Our results confirm previous findings that canonical chemokines are uniquely present in vertebrates (30, 63). Additionally, they indicate that chemokines are not evenly distributed across vertebrates. Some are very ancient, e.g., CXCL12 is present in lamprey; CXCL14 and CCL20 are present in all jawed vertebrates; and CXCL8 is present throughout bony fishes and tetrapods, with few exceptions, notably mouse and rat. However, a large part of the chemokine diversity evolved within mammals (e.g., CXCL1/2/3, CXCL16, and CCL25), particularly placentals (e.g., CXCL5/6 and CCL3/18). The phylogenetic relationships we uncovered in our reconciled tree were mostly compatible with known syntenic relationships as described in human (7). For example, the large cluster of CXC-type chemokine genes present in human chromosome 4 contains CXCL1-11 plus CXCL13 (7), all of which coalesce within a monophyletic group in our tree (Figure 5.2A). The micro-synteny within this cluster is also to some extent reflected in the phylogenetic relationships. Similarly, the other large syntenic cluster of chemokines, located on human chromosome 17, containing most of the CC-type chemokines (7), corresponds, with few exceptions, to a large monophyletic clade in our tree (Figure 5.2A). CXCL16 which is on a nearby locus of chromosome 17, is also phylogenetically related to this CC-type clade (Figure 5.2A). The complement of the canonical chemokines undergoes the largest expansion at the base of jawed vertebrates, where there is an expansion from 4 to 18 genes (Figure 5.2B). A second expansion occurred at the base of bony fishes (i.e., Osteichthyes) followed by a relative stability until placental mammals, where the total number of canonical chemokine ligands jumped to 45 genes. Finally, unlike previous works (65) our results support the presence of orthologs of both CC-type and CXC-type in the common ancestor of all vertebrates (Figure 5.2A).

Differently from the canonical chemokines, we identified a *bona fide* TAFA, i.e., with specific protein motifs, in the urochordates, the sister group to vertebrates (see Supplementary Results and Figures S6-7). The phylogenetic trees (Supplementary Figures S10,11) identified monophyletic groups for TAFA5 (TBE=0.98, UFB=98%), TAFA1 (TBE=0.94, UFB=98%), TAFA4 (TBE=0.77, UFB=75%) and TAFA2/3 (TBE=0.65, UFB=84%). The reconciled tree from GeneRax places the root at the urochordate sequence (Figure S19), therefore clarifying that the TAFA5 clade is the sister group to TAFA1-4 (Figure 5.2A). The family originated in the ancestor of urochordates and vertebrates, and the first duplications occurred at the base of vertebrates giving rise

to the TAFA5 split followed by the TAFA1 split. Subsequently, at the base of jawed vertebrates, additional duplications bring the complements from 3 to 10 (Figure 5.2B), giving rise to the remaining groups so that all jawed vertebrates possess the full diversity of TAFAs.

The phylogenetic trees for CYTL and CXCL17 mainly reflect the species trees (Figures S12-15), and the reconciliations revealed very simple complement dynamics (Figure 5.2B and Figures S20,21). However, these molecules show a remarkable difference in their distribution. CYTLs are present throughout gnathostomes, while CXCL17 is found only in placental mammals (Figure 5.2A).

The phylogenetic analysis for the CKLF super family (Figure 5.1D and Figures S16,17) recovered a monophyletic clade for the CKLF I group (TBE=0.96, UFB=80%) that we had already identified through CLANS. This group contains CKLF, that is known to interact with C-C chemokine receptor 4 (10, 11), as well as CMTM1, 2, 3, 5, and proteolipid protein 2 (PLP2). Other monophyletic clades that are consistent with the CLANS are CMTM4/6 (TBE=0.90, UFB=61%), CMTM7 (TBE=0.92, UFB=83%) and a clade containing CMTM8 plus other related molecules such as plasmolipin (PLLP) and myelin and lymphocyte proteins (MAL) (TBE=0.89, UFB=60%). The latter were all part of a large cluster that we called CKLF II in the CLANS (Figure 5.1B). However, the placement of the root of the tree in Figure 5.1D can affect the interpretation of the relationships among CKLF II subgroups. To address this problem and clarify the patterns of duplications and the presence/absence of each group throughout animals, we used GeneRax to reconcile the gene with the species tree (see above and Material and Methods for details). Our results suggest (Figure 5.2 and Figure S22) that most CKLFSF groups, such as CMTM4,6 and 8, originate in the vertebrate stem group from pre-existing CMTM genes and are widely distributed in animals. The CKLF I subgroups originate from duplications at the base of jawed vertebrates, except for the split between CKLF and CMTM1 that occurs only within mammals (Figure 5.2A). We observe the major two expansions of the CKLFSF genes in the stem group of vertebrates (from 6 to 10 complements), and then in jawed vertebrates (from 10 to 16 complements). Interestingly the extents of these expansions are less drastic than those we see for canonical chemokines (Figure 5.2B). In total, we have identified that the five distinct ligand groups have a different origin in the animal tree of life and underwent divergent evolutionary histories.

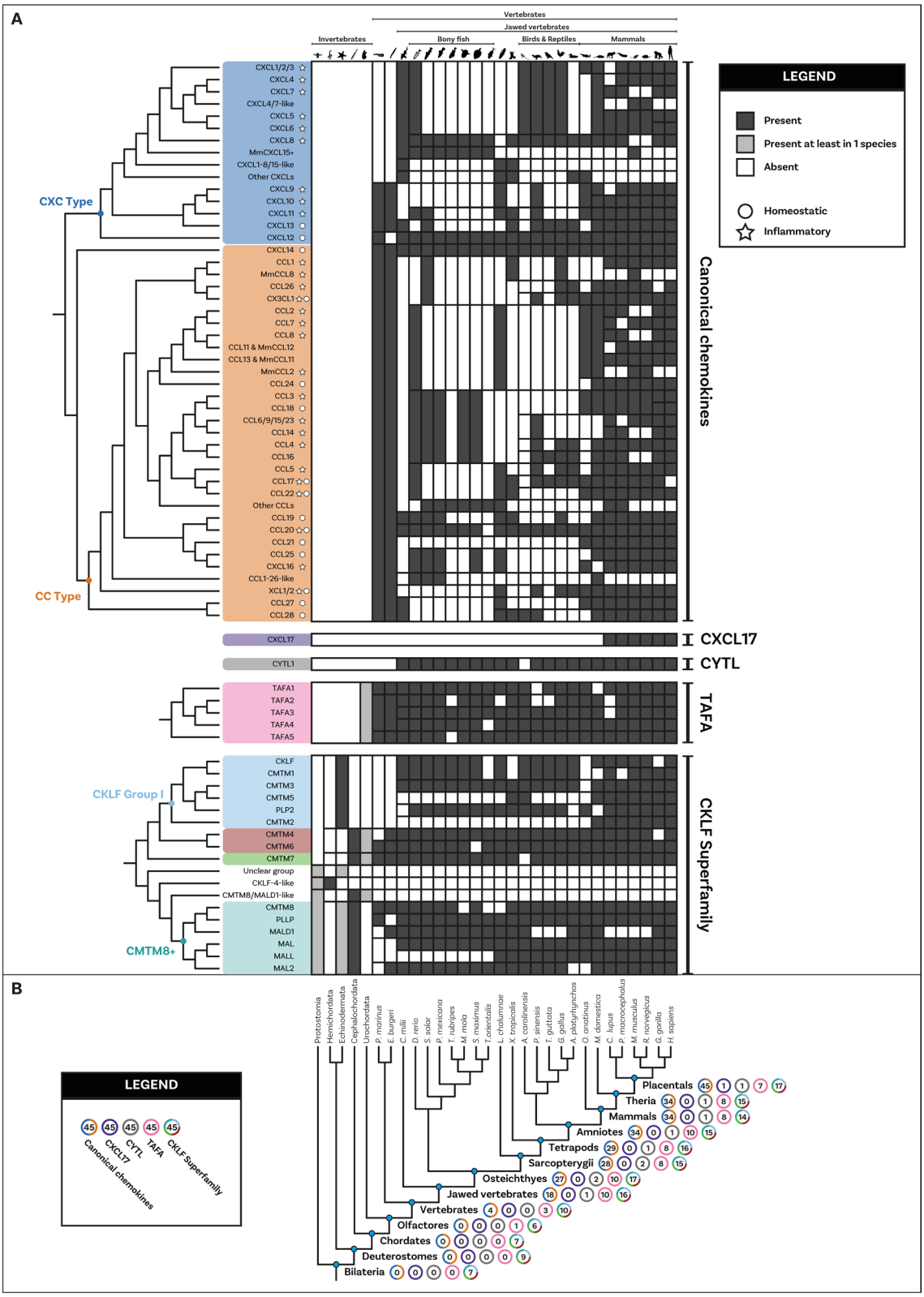

Figure 5.2. Distribution and duplication patterns of ligand groups. (A) Presence of all ligand groups are mapped onto a species tree. Gene trees and duplication events are based on the gene tree to species tree reconciliation analyses. The nomenclature for canonical chemokines is primarily based on known 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 (7)

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. Silhouette images are by Andreas Hejnol (*Xenopus laevis*); Andy Wilson (*Anas platyrhynchos*, *Taeniopygia guttata*); Carlos Cano-Barbacil (*Salmo trutta*); Christoph Schomburg (*Anolis carolinensis*, *Ciona intestinalis*, *Eptatretus burgeri*, *Petromyzon marinus*); Christopher Kenaley (*Mola mola*); Chuanixn Yu (*Latimeria chalumnae*); Daniel Jaron (*Mus musculus*); Daniel Stadtmauer (*Monodelphis domestica*); Fernando Carezzano (Asteroidea); Ingo Braasch (*Callorhinchus milii*); Jake Warner (*Danio rerio*); Kamil S. Jaron (*Poecilia formosa*); Mali'o Kodis, photograph by Hans Hillewaert (*Branchiostoma lanceolatum*, https://www.phylopic.org/images/719d7b41-cedc-4c97-9ffe-dd8809f85553/branchiostoma-lanceolatum); Margot Michaud (*Canis lupus*, *Physeter macrocephalus*); NASA (*Homo sapiens sapiens*); Nathan Hermann (*Scophthalmus aquosus*); Ryan Cupo (*Rattus norvegicus*); seung9park (*Takifugu rubripes rubripes*); Soledad Miranda-Rottmann (*Pelodiscus sinensis*, https://www.phylopic.org/images/929fd134-bbd7-4744-987f-1975107029f5/pelodiscus-sinensis); Steven Traver (*Gallus gallus domesticus*, *Ornithorhynchus anatinus*); Stuart Humphries (*Thunnus thynnus*); T. Michael Keesey (after Colin M. L. Burnett) (*Gorilla gorilla gorilla*); Thomas Hegna (based on picture by Nicolas Gompel) (*Drosophila (Drosophila) mojavensis*); and Yan Wong (*Balanoglossus*).

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

Next, we investigated the origin and pattern of duplication for the chemokine receptors and chemokine-like receptors (Table 5.1). Using BLASTP against the 64 species, we identified 7,157 putative chemokine receptors (see Materials and Methods for more details), and we investigated their relationships using CLANS (see above for justification). The result (Figure S23C) identifies four main groups of chemokine receptors and chemokine-like receptors. The first comprises canonical receptors (i.e., CCR, CXCR, CX3CR1, CX3C, and XCR1), and the second includes atypical receptor 3 and GPR182, which has been recently shown to have chemokine receptor activity (66). The third group, which we named Chemokine-like plus (CML-plus), contains the chemokine-like receptors (CML1 also known as chemerin receptor 1), formyl peptide receptors (FPR) that bind the TAFA ligands (15, 16) and other GPCRs such as GPR1 (chemerin receptor 2), GPR33, PTGDR2. Furthermore, the CLANS analysis identifies an intermediate group containing angiotensin, apelin and other receptors and shows sequence similarity to canonical and chemokine-like receptors (Figure S23B). Finally,

our analysis identifies a small cluster composed of only ACKR1 that do not connect to other GPCRs or other atypical receptors even at loose p-value thresholds. This indicates that their sequence is either non-homologous or highly divergent from other chemokine receptors and atypical receptors. Overall, these 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/GPR182. However, the results indicate that the other GPCRs, such as the chemokine-like receptors, formyl peptide receptors, GPR1, and GPR33, 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 (67) or adrenomedullin receptors (30, 68). However, these works were based on error-prone phylogenetic methods such as Neighbour Joining (68). 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 receptor's outgroup. To clarify this, we lowered the p-value thresholds of CLANS (to p-value < $1e\text{-}^{50}$) and collected a combined dataset including all chemokine receptor sequences and outgroups (i.e., sequences that connect to the chemokine receptor 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 chemokine receptors (UFB=96, TBE=1.0), the CML-plus (UFB=95, TBE=0.99) and the atypical 3/GPR182 (UFB=100, TBE=1) (Figure 5.3, S24 and S25). In contrast, viral chemokine receptors are paraphyletic, with three sequences placed

within the canonical chemokine receptors 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., GPR25; GPR15) forms a monophyletic clade with the canonical chemokine receptors, CML-plus group and atypical3/GPR182 (UFB=61, TBE=0.91). However, its position changes between the sister group to canonical chemokine receptors plus atypical3/GPR182 in the TBE tree (TBE=0.84) and sister to CML plus in the ultrafast bootstrap tree (UFB=38).

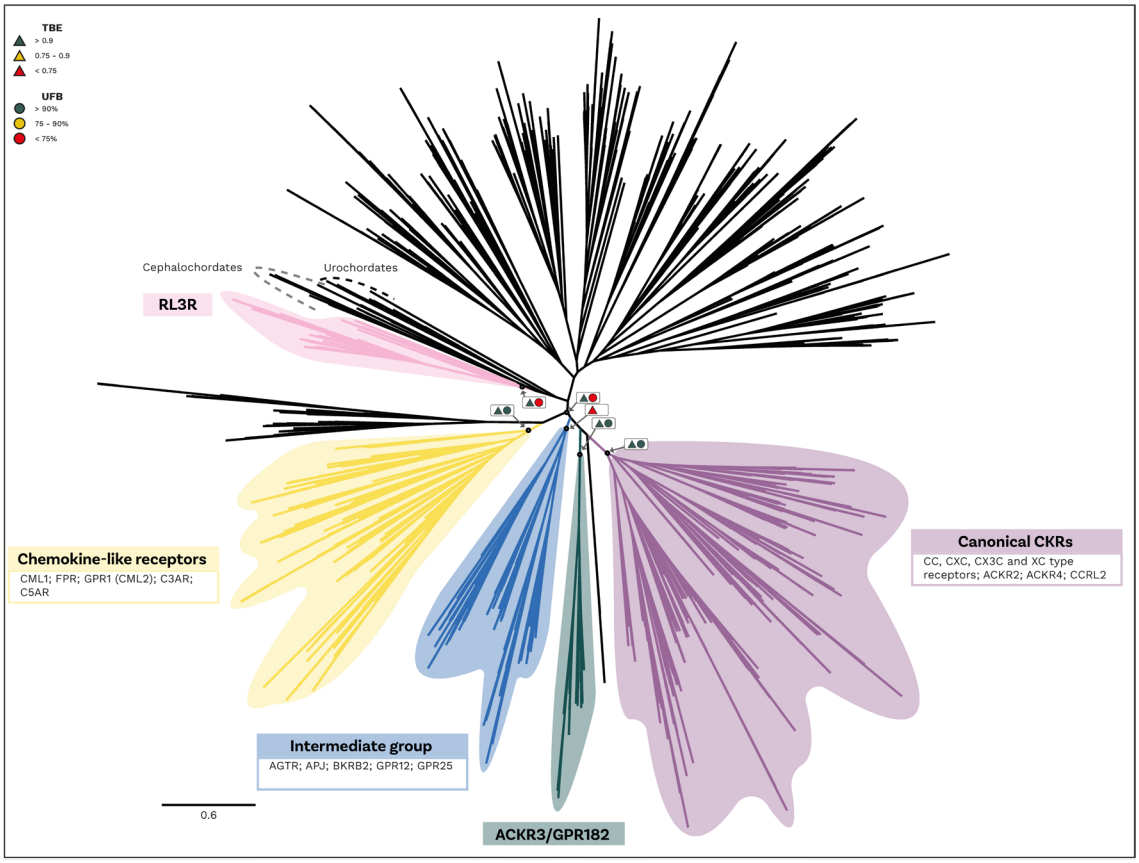

Figure 5.3. Phylogeny of Receptor groups. An unrooted maximum likelihood phylogeny of chemokine receptors. The tree shown is the transfer bootstrap expectation (TBE) tree including just the chordate specific clade from the ultrafast bootstrap tree (UFB). Node supports from both TBE (triangle) and UFB (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.

All the groups mentioned above form a large clade composed of vertebrate-specific GPCRs (UBF=100 TBE=0.96) that also includes other GPCRs, such as CLTR and P2RY receptors (Figure 5.3, S24 and S25). Another orphan GPCR, GPR35, had been

proposed as a potential chemokine receptor (50); however, this was later questioned (56, 57) and GPR35 is still generally considered orphan (69–71). Our analysis collected GPR35 and placed it within this large vertebrate specific clade indicating that is it also a vertebrate specific gene but not phylogenetically a 'canonical' chemokine receptor. The closest outgroup to this clade is composed of a few sequences from urochordates, the sister group of vertebrates (UFB=49 TBE=0.91) (Figure 5.3, S24 and S25). Interestingly, as the sister group of this clade, we identify a group composed of Relaxin receptors which contain sequences from both urochordates and vertebrates (UBF=53 TBE=0.95). Finally, as the sister group of these large clades, we identified a clade of cephalochordate-specific sequences (UBF=44).

To clarify the duplication pattern and origin of the chemokine receptors, we used GeneRax (61) (see Materials and Methods). Our results indicate (Figure 5.4A, S26) that all chemokine receptors (except ACKR1) originated from a duplication in the stem lineage of vertebrates. This duplication of an unknown GPCR gave rise to the CML-plus, the canonical chemokine receptors and atypical 3/GPR182 groups as well as the intermediate group and other GPCRs (Figure 5.4A, S26). This result is consistent with the distribution of the paralogous Relaxin receptors which are present both in urochordates and vertebrates and the position of the orphan urochordate sequences as sister group of canonical chemokine receptors, CML-plus group and atypical3/GPR182 and other GPCRs (see above). Furthermore, the phylogenetic relationships amongst canonical chemokine receptors are overall consistent with the syntenic gene patterns known in human (7). The largest cluster of chemokine receptor genes spans 3 closely located loci on human chromosome 3 (7). It includes most CCRs as well as XCR and CX3CR and corresponds to one of the two major monophyletic clades in our tree (Figure 5.4A). Another example is the mini cluster of CXCR1 and CXCR2, located on human chromosome 2 (7) that we also found to form a monophyletic clade (Figure 5.4A).

We used the reconciliation to better understand the repertoires of receptors present at key nodes during vertebrate evolution. Our results (Figure 5.4B) show a substantial difference in the duplication pattern of different receptor families. For example, the complement of the atypical3/GPR182 remains constant throughout vertebrate evolution while the canonical and chemokine-like receptor groups expanded dramatically. The canonical chemokine receptors expanded from 5 to 20 genes and the CML-plus from 1 to 11 in the ancestor of the jawed vertebrates (Figure 5.4B). 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 receptor 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 GPR182 as well as a single copy gene which would later diverge to produce all the CML-plus clade.

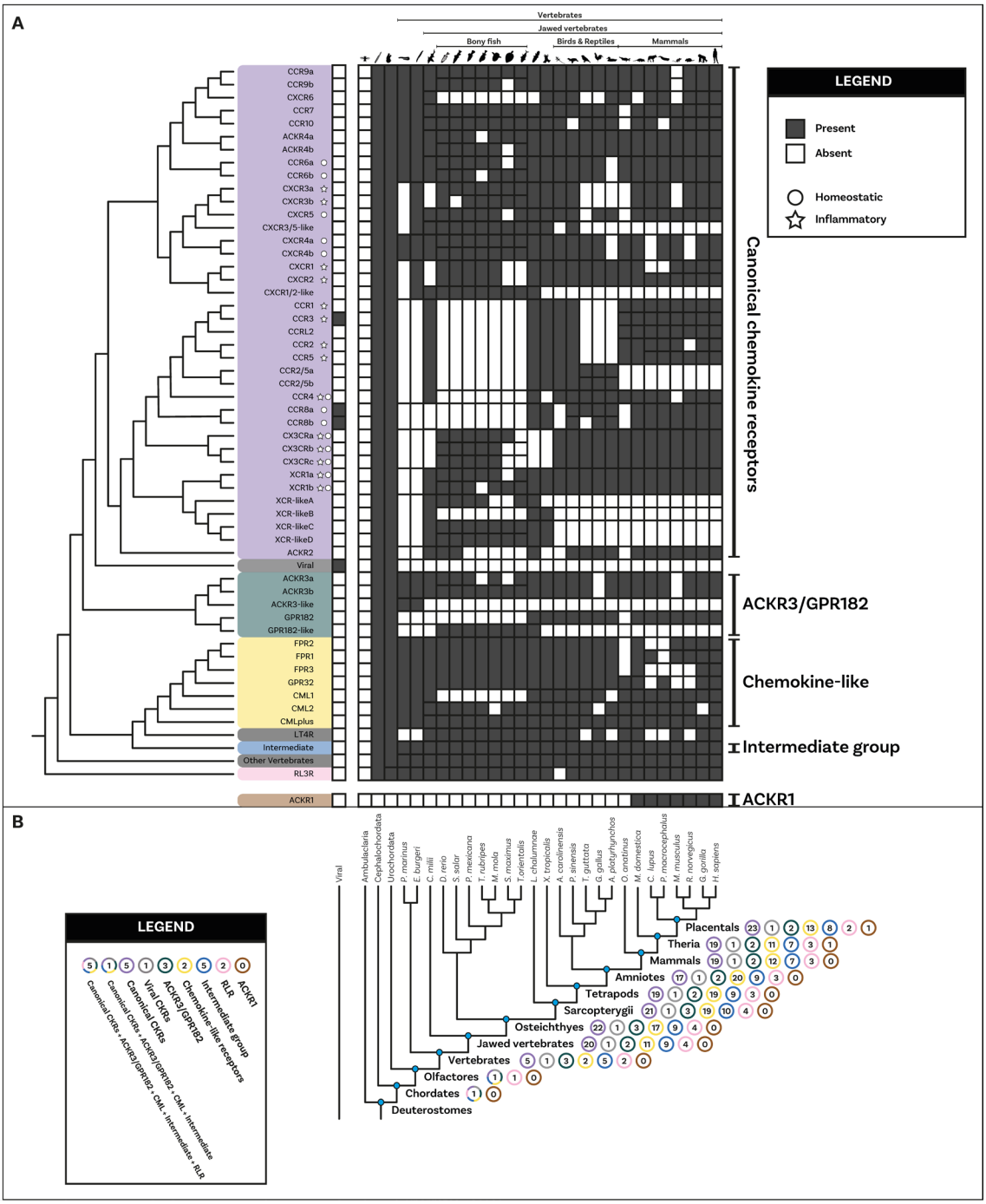

Figure 5.4. Distribution and duplication patterns of receptor groups. (A) Presence of all receptor groups are mapped onto a species tree. Gene trees and duplication events are based on the gene tree to

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 (7). **(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. Silhouette images are by Andreas Hejnol (*Xenopus laevis*); Andy Wilson (*Anas platyrhynchos*, *Taeniopygia guttata*); Carlos Cano-Barbacil (*Salmo trutta*); Christoph Schomburg (*Anolis carolinensis*, *Ciona intestinalis*, *Eptatretus burgeri*, *Petromyzon marinus*); Christopher Kenaley (*Mola mola*); Chuanixn Yu (*Latimeria chalumnae*); Daniel Jaron (*Mus musculus*); Daniel Stadtmauer (*Monodelphis domestica*); Fernando Carezzano (Asteroidea); Ingo Braasch (*Callorhinchus milii*); Jake Warner (*Danio rerio*); Kamil S. Jaron (*Poecilia formosa*); Mali'o Kodis, photograph by Hans Hillewaert (*Branchiostoma lanceolatum*, https://www.phylopic.org/images/719d7b41-cedc-4c97-9ffe-dd8809f85553/branchiostoma-lanceolatum); Margot Michaud (*Canis lupus*, *Physeter macrocephalus*); NASA (*Homo sapiens sapiens*); Nathan Hermann (*Scophthalmus aquosus*); Ryan Cupo (*Rattus norvegicus*); seung9park (*Takifugu rubripes rubripes*); Soledad Miranda-Rottmann (*Pelodiscus sinensis*, https://www.phylopic.org/images/929fd134-bbd7-4744-987f-1975107029f5/pelodiscus-sinensis); Steven Traver (*Gallus gallus domesticus*, *Ornithorhynchus anatinus*); Stuart Humphries (*Thunnus thynnus*); T. Michael Keesey (after Colin M. L. Burnett) (*Gorilla gorilla gorilla*); Thomas Hegna (based on picture by Nicolas Gompel) (*Drosophila (Drosophila) mojavensis*); and Yan Wong (*Balanoglossus*).

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 5.5) suggest that the key components of the chemokine system, including the chemokine receptors themselves, evolved in the stem group of vertebrates in the Cambrian around 500 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.

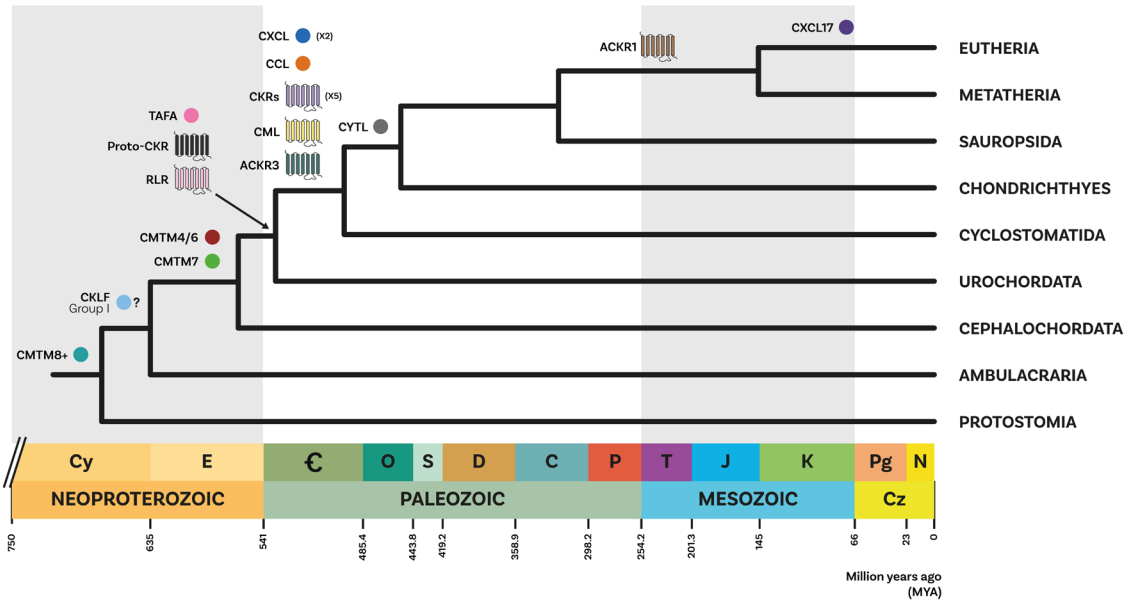

Figure 5.5. Summary of the evolution of ligands and receptors. A summary diagram of the evolution of the different chemokine system components. A simplified phylogenetic tree of species is shown, calibrated to time according to Dohrmann and Wörheide 2017 (72) for Deuterostomia and Bilateria nodes and Delsuc et al. 2018 (73) for all other nodes. Circles represent ligand groups, and 7 transmembrane domain structure icons represent 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 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 5.2). Geological column is shown along the bottom, in accordance with the ICS International Chronostratigraphic Chart (74).

Unrelated molecules converged to chemokine function.

Based on the presence of shared protein motifs, TAFA "chemokines" (13, 14), CXCL17 (75, 76) and CYTL (19) have been proposed to be homologous to chemokine ligands. However, our findings strongly suggest that these molecules are not homologous (Figure 5.1) and likely acquired the ability to activate a chemokine-like response through convergent evolution. Our conclusions differ from those previous studies (13, 19, 75, 76) 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 (29), 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 (12) 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 5.2) and confirm that they are present only in vertebrates (30). We identify orthologs of CXCL and CCL ligands in both extant lineages of cyclostomes (Figure 5.2A). While chemokines have already been described in lamprey (65, 77, 78), 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 controversial putative receptor, GPR35 (50, 56, 57), 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 considered part of the CKLF super family are closely related to classic members and should be included in it. For example, proteolipid protein 2 (PLP2) belongs to the CKLF I group and is, therefore more closely related to the CKLF with chemokine function than several other CKLFSF members. Similarly, CMTM8 is more closely related to plasmolipin (PLLP) and myelin and lymphocyte protein (MAL) than to any of the classic CKLFSF members. Although this relationship had been proposed based only on sequence similarity (34), our phylogenetic analysis provides additional evidence for it. Therefore, the potential chemokine function of all these additional members should be explored *in vitro* and *in vivo* in both vertebrates and invertebrates.

All receptors but one derive from a single gene duplication.

Our results clarify the distribution of canonical chemokine receptors in vertebrates (Figure 5.4), and their evolutionary relationships and identify the pattern of duplication that leads to their origin (Figure 5.4A, S26). Unlike previous works (79), we identify that atypical receptors do not form a monophyletic group. Specifically, atypical 2 and 4 are part of the canonical clade specifically related to CC-type receptor subclades. Furthermore, we find that the atypical 3 receptors are related to GPR182, supporting previous functional data suggesting that the latter are atypical chemokine receptors binding CXCL10, 12, and 13 (66). We attribute these differences to our use of wider GPCR sampling and improved methods for phylogenetic inference.

Remarkably, our results do not identify ACKR1 as related to the main chemokine receptors but rather as a divergent clade (Figure S23). To our knowledge, this is the first time this observation has been made. Our current results do not allow us to clarify the evolutionary origin of ACKR1. However, the presence of 7TMD domains suggests that they are GPCRs that independently acquired the ability to bind chemokines.

Alternatively, similarly to other genes evolved in the immune system, ACKR1 may have been subjected to strong selective pressures that substantially changed their sequence, obscuring their phylogenetic relationships. The case of ACKR1 being the most distantly related receptor is intriguing as it is one of the most promiscuous chemokine receptors (2, 80) and it has been shown to bind both CC and CXC chemokines (81, 82).

Viral chemokine receptors represent a cryptic group that can bind multiple chemokines (22, 23). Despite their functional similarity to canonical chemokine receptors, viral chemokine receptors' 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. The placement of viral sourced sequences within an otherwise vertebrate specific clade supports the hypothesis that viruses acquired these genes through non-vertical inheritance. Given the paraphyly of viral receptors, this appears to have occurred multiple times. However, there are significant uncertainties and further work is needed to untangle details of viral chemokine receptors' evolution.

Our analysis reveals that the clade comprising apelin receptors, angiotensin receptors, bradykinin receptors, and orphan GPCRs (shown in Figure 5.3, 5.4 and S24–26) is closely related to chemokine receptors. This finding partially supports previous studies (67) that suggested a gene duplication event gave rise to both chemokine receptors 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, GPR182, chemokine-like receptors, formyl peptide receptors, the intermediate group, and many other known and orphan GPCRs including the controversial putative CXCL17 receptor GPR35. These findings suggest that two rounds of genome duplication (83, 84) 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 and produced a comprehensive description of the distribution of these molecules throughout animals (Figures 5.2 and 5.4). Chemokine and chemokine receptor repertoires are known to vary even amongst closely related species (85). Moreover, technical difficulties in identifying true homologs when working with fast evolving short sequences pose additional challenges to study chemokine evolution. Despite this, our broad and diverse species sampling has allowed us to elucidate the evolutionary history of these molecules with considerable detail. While we cannot exclude that some absences may arise as artifacts (sequences may remain undetected for instance due to stringent BLAST e-value thresholds for highly diverged sequences or due to incomplete genomes/proteomes), overall, we were able to trace the presence/absence of major groups of chemokine components throughout animals. Our analysis suggests that the canonical chemokine signaling evolved in the vertebrate stem group (about 500 Mya) likely due to the two rounds of genome duplication that gave rise to many vertebrate novelties (83, 84). 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 GPR182 (Figure 5.5). 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 (7). Chemokine ligand and receptor genes are known to cluster on specific chromosomes (7) consistent with the hypothesis that they may be the result of the combination of *en bloc* duplication followed by tandem duplications (30, 63, 64). Due to limited high-quality genomes, syntenic patterns of chemokine genes described so far are based primarily on human and a handful of other species (30, 63, 64), hampering our grasp of the level of conservation of these syntenic patterns. Conversely, our large-scale phylogenetic analyses encompassed many species. We uncovered several phylogenetic relationships that are consistent with known syntenic patterns in human, providing stronger evidence

for their evolutionary relationship. Minor discrepancies between phylogenetic relationships and syntenic patterns are interesting source of future investigation into the conservation of syntenic patterns throughout vertebrate history, as high-quality genomic data become more widely available.

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 (7). 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 (86) - a key vertebrate innovation (87). This combined with the fact that homeostatic chemokine ligands/receptors tend to be restricted to monogamous pairing (2, 85) suggests that homeostatic chemokine pairings are more ancient and conserved being in single copy throughout much of the vertebrates. Contrastingly, inflammatory chemokine pairings are more promiscuous, and this could be linked to the more recent duplications in the genes, such as for CCL2/7/8/11/13 (Figure 5.2A) and their receptors CCR1/2/3/4/5 (Figure 5.4A). 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 5.5), 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 S1). BUSCO v4.0.6 (88, 89) and the metazoa_odb10 set of 954 genes were used to evaluate their completeness (Table S1).

To identify potential homologs of canonical chemokines, TAFA chemokines and CYTL1, we used 207 curated sequences that we obtained from SwissProt (90, 91) as seeds for an initial BLASTP (51, 53) with e-value < e^{-10}. To identify putative chemokines in cyclostomes, the lamprey *Petromyzon marinus* (92) and the hagfish *Eptatretus burger* (93), 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 homologs spanning multiple invertebrate phyla. Further characterisation of these invertebrate sequences, through BLASTP versus SwissProt, protein domains search with InterProScan (94, 95), position in CLANS analysis (see below) and, where necessary, multiple sequence alignments, led us to retain only one urochordate sequence as a putative TAFA homolog (see Supplementary Results for details).

To identify homologs for the CKLF superfamily, we used 21 SwissProt-reviewed sequences. In addition to the BLASTP search, we used a position-specific iterative BLAST (PSI-BLAST) (52) with an e-value threshold of <e^{-10}. Using this approach, we identified a total of 590 putative homologs, including 186 from invertebrates.

We used BLASTP using 178 manually annotated receptor sequences from SwissProt as query sequences for the chemokine receptors. This includes all human canonical and atypical chemokine receptors (96). We also collected 8 viral sequences with chemokine receptor activity from UniProt (97) and performed a second BLASTP. We extracted all BLAST hits with e-values < e^{-10} and used Phobius (98) to predict their transmembrane domain structure. Only sequences with 5-8 transmembrane domains were kept. Hit sequences were annotated by their top 5 BLAST hits against SwissProt. All hits from both BLASTs were merged and filtered by cd-hit (99, 100) to remove redundant

sequences at the 95% similarity threshold. This resulted in 7,157 putative chemokine GPCR sequences.

Identification of subgroups with Cluster Analysis of Sequences (CLANS).

We utilized CLANS (54, 55) with default parameters and different p-values (i.e., stringency values) to visualize the relationships between subgroups of ligands and receptors. We assessed the similarity and interrelationships between different clusters by gradually relaxing the p-value threshold (Figures S1, S2 and S23). Additionally, we annotated each cluster using gene annotations for key species *Homo*, *Mus*, *Gorilla*, *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 (101) were added to the dataset as well as the 8 seed viral chemokine receptors from UniProt (97).

Alignment and phylogenetic analysis.

Alignment.

All ligand and receptor sequences were aligned using MAFFT (102, 103) with the --auto setting and using trimAl (104) to remove positions with >70% gaps.

Gene Trees.

All gene alignments were analysed using IQTREE2 (105), the model test algorithm (106) was used to select the best substitution model for each analysis. Best models selected by IQTREE2 for each set are listed in Table S2 (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 (58, 59) 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 (60).

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 receptor 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 (61). 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 (107) 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. K. Zhang, S. Shi, W. Han, Research progress in cytokines with chemokine-like function. *Cellular & Molecular Immunology* **15**, 660–662 (2018).
2. K. Chen, *et al.*, Chemokines in homeostasis and diseases. *Cell Mol Immunol* **15**, 324–334 (2018).
3. X. Blanchet, M. Langer, C. Weber, R. Koenen, P. von Hundelshausen, Touch of Chemokines. *Frontiers in Immunology* **3**, 175 (2012).
4. 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).
5. P. B. Tran, R. J. Miller, Chemokine receptors: signposts to brain development and disease. *Nature Reviews Neuroscience* **4**, 444–455 (2003).
6. B. Moser, M. Wolf, A. Walz, P. Loetscher, Chemokines: multiple levels of leukocyte migration control☆. *Trends in Immunology* **25**, 75–84 (2004).
7. A. Zlotnik, O. Yoshie, The Chemokine Superfamily Revisited. *Immunity* **36**, 705–716 (2012).

8. A. Zlotnik, O. Yoshie, Chemokines: A New Classification System and Their Role in Immunity. *Immunity* **12**, 121–127 (2000).
9. H. Nomiyama, N. Osada, O. Yoshie, A family tree of vertebrate chemokine receptors for a unified nomenclature. *Developmental & Comparative Immunology* **35**, 705–715 (2011).
10. Y. Wang, *et al.*, Chemokine-like factor 1 is a functional ligand for CC chemokine receptor 4 (CCR4). *Life Sciences* **78**, 614–621 (2006).
11. Y. Wang, *et al.*, Two C-terminal peptides of human CKLF1 interact with the chemokine receptor CCR4. *The International Journal of Biochemistry & Cell Biology* **40**, 909–919 (2008).
12. D.-D. Liu, *et al.*, Progress in pharmacological research of chemokine like factor 1 (CKLF1). *Cytokine* **102**, 41–50 (2018).
13. Y. Tom Tang, *et al.*, TAFA: a novel secreted family with conserved cysteine residues and restricted expression in the brain. *Genomics* **83**, 727–734 (2004).
14. D. C. Sarver, X. Lei, G. W. Wong, FAM19A (TAFA): An Emerging Family of Neurokines with Diverse Functions in the Central and Peripheral Nervous System. *ACS Chem. Neurosci.* **12**, 945–958 (2021).
15. W. Wang, *et al.*, FAM19A4 is a novel cytokine ligand of formyl peptide receptor 1 (FPR1) and is able to promote the migration and phagocytosis of macrophages. *Cell Mol Immunol* **12**, 615–624 (2015).
16. M. Y. Park, *et al.*, FAM19A5, a brain-specific chemokine, inhibits RANKL-induced osteoclast formation through formyl peptide receptor 2. *Sci Rep* **7**, 15575 (2017).
17. C. Zheng, *et al.*, FAM19A1 is a new ligand for GPR1 that modulates neural stem-cell proliferation and differentiation. *The FASEB Journal* **32**, 5874–5890 (2018).
18. X. Wang, *et al.*, Cytokine-like 1 Chemoattracts Monocytes/Macrophages via CCR2. *The Journal of Immunology* **196**, 4090–4099 (2016).
19. A. Tomczak, M. T. Pisabarro, Identification of CCR2-binding features in Cytl1 by a CCL2-like chemokine model. *Proteins: Structure, Function, and Bioinformatics* **79**, 1277–1292 (2011).
20. T. Yoshimura, J. J. Oppenheim, Chemokine-like receptor 1 (CMKLR1) and chemokine (C–C motif) receptor-like 2 (CCRL2); Two multifunctional receptors with unusual properties. *Experimental Cell Research* **317**, 674–684 (2011).

21. R. Bonecchi, G. J. Graham, Atypical Chemokine Receptors and Their Roles in the Resolution of the Inflammatory Response. *Frontiers in Immunology* **7**, 224 (2016).
22. T. N. Kledal, M. M. Rosenkilde, T. W. Schwartz, Selective recognition of the membrane-bound CX3C chemokine, fractalkine, by the human cytomegalovirus-encoded broad-spectrum receptor US28. *FEBS Letters* **441**, 209–214 (1998).
23. T. F. Miles, *et al.*, Viral GPCR US28 can signal in response to chemokine agonists of nearly unlimited structural degeneracy. *eLife* **7**, e35850 (2018).
24. M. M. Rosenkilde, M. J. Smit, M. Waldhoer, Structure, function and physiological consequences of virally encoded chemokine seven transmembrane receptors. *British Journal of Pharmacology* **153**, S154–S166 (2008).
25. H. Daiyasu, W. Nemoto, H. Toh, Evolutionary Analysis of Functional Divergence among Chemokine Receptors, Decoy Receptors, and Viral Receptors. *Frontiers in Microbiology* **3** (2012).
26. M. Meyrath, *et al.*, The atypical chemokine receptor ACKR3/CXCR7 is a broad-spectrum scavenger for opioid peptides. *Nature Communications* **11**, 3033 (2020).
27. R. J. B. Nibbs, G. J. Graham, Immune regulation by atypical chemokine receptors. *Nat Rev Immunol* **13**, 815–829 (2013).
28. F. Bachelerie, *et al.*, New nomenclature for atypical chemokine receptors. *Nat Immunol* **15**, 207–208 (2014).
29. S. S. Denisov, CXCL17: The Black Sheep in the Chemokine Flock. *Frontiers in Immunology* **12**, 2811 (2021).
30. M. E. DeVries, *et al.*, Defining the Origins and Evolution of the Chemokine/Chemokine Receptor System. *The Journal of Immunology* **176**, 401 (2006).
31. B. Bajoghli, Evolution and function of chemokine receptors in the immune system of lower vertebrates. *European Journal of Immunology* **43**, 1686–1692 (2013).
32. 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).
33. 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).

34. 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).
35. 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).
36. W. Han, *et al.*, Molecular cloning and characterization of chemokine-like factor 1 (CKLF1), a novel human cytokine with unique structure and potential chemotactic activity. *Biochem J* **357**, 127–135 (2001).
37. L. Wang, *et al.*, Molecular cloning and characterization of chemokine-like factor super family member 1 (CKLFSF1), a novel human gene with at least 23 alternative splicing isoforms in testis tissue. *The International Journal of Biochemistry & Cell Biology* **36**, 1492–1501 (2004).
38. C. Jin, P. Ding, Y. Wang, D. Ma, Regulation of EGF receptor signaling by the MARVEL domain-containing protein CKLFSF8. *FEBS Letters* **579**, 6375–6382 (2005).
39. Z.-Z. Wang, *et al.*, Chemokine-like factor 1, a novel cytokine, induces nerve cell migration through the non-extracellular Ca2+-dependent tyrosine kinases pathway. *Brain Research* **1308**, 24–34 (2010).
40. T. Li, *et al.*, Expression of chemokine-like factor 1 is upregulated during T lymphocyte activation. *Life Sciences* **79**, 519–524 (2006).
41. Y. Zhang, *et al.*, C-terminal peptides of chemokine-like factor 1 signal through chemokine receptor CCR4 to cross-desensitize the CXCR4. *Biochemical and Biophysical Research Communications* **409**, 356–361 (2011).
42. H. Li, *et al.*, A novel 3p22.3 gene CMTM7 represses oncogenic EGFR signaling and inhibits cancer cell growth. *Oncogene* (2014) https:/doi.org/10.1038/onc.2013.282.
43. S. Zhu, *et al.*, Protein Cytl1: its role in chondrogenesis, cartilage homeostasis, and disease. *Cell. Mol. Life Sci.* **76**, 3515–3523 (2019).
44. H. Xue, *et al.*, CYTL1 Promotes the Activation of Neutrophils in a Sepsis Model. *Inflammation* **43**, 274–285 (2020).
45. X. Wang, *et al.*, Tafa-2 plays an essential role in neuronal survival and neurobiological function in mice. *Acta Biochimica et Biophysica Sinica* **50**, 984–995 (2018).

46. Y. Wang, *et al.*, Novel Adipokine, FAM19A5, Inhibits Neointima Formation After Injury Through Sphingosine-1-Phosphate Receptor 2. *Circulation* **138**, 48–63 (2018).
47. J. Okada, *et al.*, Analysis of FAM19A2/TAFA-2 function. *Physiology & Behavior* **208**, 112581 (2019).
48. B. L. Lokeshwar, G. Kallifatidis, J. J. Hoy, "Chapter One - Atypical chemokine receptors in tumor cell growth and metastasis" in *Advances in Cancer Research*, GPCR Signaling in Cancer., A. K. Shukla, Ed. (Academic Press, 2020), pp. 1–27.
49. H.-Q. He, R. D. Ye, The Formyl Peptide Receptors: Diversity of Ligands and Mechanism for Recognition. *Molecules* **22**, 455 (2017).
50. 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).
51. 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).
52. 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).
53. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
54. T. Frickey, A. Lupas, CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
55. F. Gabler, *et al.*, Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics* **72**, e108 (2020).
56. S.-J. Park, S.-J. Lee, S.-Y. Nam, D.-S. Im, GPR35 mediates lodoxamide-induced migration inhibitory response but not CXCL17-induced migration stimulatory response in THP-1 cells; is GPR35 a receptor for CXCL17? *British Journal of Pharmacology* **175**, 154–161 (2018).
57. N. A. S. B. M. Amir, *et al.*, Evidence for the Existence of a CXCL17 Receptor Distinct from GPR35. *The Journal of Immunology* **201**, 714–724 (2018).
58. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution* **30**, 1188–1195 (2013).
59. 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).

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

61. 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).

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

63. 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).

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

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

66. 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).

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

68. 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).

69. S. Xiao, W. Xie, L. Zhou, Mucosal chemokine CXCL17: What is known and not known. *Scandinavian Journal of Immunology* **93**, e12965 (2021).

70. S. P. Giblin, J. E. Pease, What defines a chemokine? – The curious case of CXCL17. *Cytokine* **168**, 156224 (2023).

71. J. Duan, *et al.*, Insights into divalent cation regulation and G13-coupling of orphan receptor GPR35. *Cell Discov* **8**, 1–12 (2022).

72. M. Dohrmann, G. Wörheide, Dating early animal evolution using phylogenomic data. *Sci Rep* **7**, 3599 (2017).

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

74. F. M. Gradstein, J. G. Ogg, “Chapter 2 - The Chronostratigraphic Scale” in *The Geologic Time Scale*, F. M. Gradstein, J. G. Ogg, M. D. Schmitz, G. M. Ogg, Eds. (Elsevier, 2012), pp. 31–42.

75. 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).

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

77. 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).

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

79. 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).

80. S. J. Allen, S. E. Crown, T. M. Handel, Chemokine:Receptor Structure, Interactions, and Antagonism. *Annual Review of Immunology* **25**, 787–820 (2007).

81. R. Horuk, *et al.*, A Receptor for the Malarial Parasite Plasmodium vivax: the Erythrocyte Chemokine Receptor. *Science* **261**, 1182–1184 (1993).

82. R. Horuk, The Duffy Antigen Receptor for Chemokines DARC/ACKR1. *Frontiers in Immunology* **6** (2015).

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

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

85. P. M. Murphy, “15 - Chemokines and Chemokine Receptors” in *Clinical Immunology (Sixth Edition)*, R. R. Rich, *et al.*, Eds. (Elsevier, 2023), pp. 215–227.

86. 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).

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

88. R. M. Waterhouse, *et al.*, BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* **35**, 543–548 (2018).
89. 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).
90. 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.
91. S. Poux, *et al.*, On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* **33**, 3454–3460 (2017).
92. J. J. Smith, *et al.*, Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat Genet* **45**, 415–421 (2013).
93. D. Yu, *et al.*, Hagfish genome illuminates vertebrate whole genome duplications and their evolutionary consequences. 2023.04.08.536076 (2023).
94. E. M. Zdobnov, R. Apweiler, InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
95. P. Jones, *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
96. 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).
97. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).
98. 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).
99. W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
100. 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).

101. 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).
102. 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).
103. 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).
104. 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).
105. 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).
106. 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).
107. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* **33**, 1635–1638 (2016).