The evolution of signalling systems in animals: insights from vision and chemokines

Thesis submitted for the degree of Doctor of Philosophy at The University of Leicester

By

Alessandra Aleotti

Department of Genetics and Genome Biology

College of Life Sciences

University of Leicester

October 2023

Abstract

The evolution of animals from unicellular ancestors to obligate multicellular organisms has deeply influenced their biology. Integral to this transformation is the development of distinct cell types that, while fulfilling specialized roles, must communicate and coordinate to uniformly respond to internal and external stimuli. Consequently, cell signalling systems are of fundamental importance in animals. This thesis delves into the evolution of two prominent biological processes in animals underpinned by cell signalling: vision and the chemokine system. Vision, widespread in animals, has origins in the early stages of animal evolution. It relies on a photosensitive molecule, the opsin bound to a vitamin A derivative, that when hit by light, triggers a phototransduction pathway within photoreceptor cells. The phototransduction pathway involves a diverse suite of molecular components whose evolutionary journey is meticulously explored, as well as the regulatory genes involved in photoreceptor cell identity, in Chapter 3. While the evolutionary history of enzymes involved in the vitamin A metabolism is explored in Chapter 4. The chemokine system is instrumental in directing cell migration during immunity, homeostasis, and development in vertebrate. This system is composed of ligand and receptor components, with both essential "canonical" elements and additional "non-canonical" elements that interact with the system. This intricate molecular diversity is investigated in Chapter 5. Using comprehensive phylogenetic and bioinformatic methodologies, the evolution of these signalling systems is dissected. The research paints a detailed evolutionary picture of each molecular component in both systems. A recurrent theme emerged, highlighting the significance of gene family expansions at critical species nodes.

Acknowledgements

I would like to thank my supervisors: Roberto Feuda for allowing me to conduct my PhD project in his research group and for introducing me to the world of phylogenetics and bioinformatics; and Flaviano Giorgini for his support and for providing useful inputs both scientifically and professionally. I would also like to extend my gratitude to the whole Neurogenetics Group, as the lively scientific environment has helped me to learn a lot and grow scientifically.

I am extremely grateful to have been accompanied throughout these four years by an awesome cohort of fellow PhD students, both in pre- and post-pandemic times. Thanks to all of you for having made these years more pleasant.

Huge thank you goes to my little “siblings”—Clifton, Matt, Julien, DaeNia and Frannie. Without you guys around I seriously don’t know if I could have made it. Your moral and practical support was invaluable.

I would like to acknowledge my family and close friends from Italy for their encouragement throughout the years. Particularly, my sister Francesca and my parents, Luca and Silvia, for always going above and beyond to support me and for cheering me on.

Luis, you were by my side every step of the way. I would be lost without you. Thank you for your guidance and support—both in science and in life. Most importantly, thank you for cooking me delicious food every day throughout my thesis-writing journey.

Table of Contents

Chapter 1

General Introduction

General Introduction

The origin of multicellularity: a major evolutionary transition

The astonishing diversity of life on Earth showcases the profound impact of evolution over billions of years. Underpinning the complexity of life forms are major evolutionary transitions, landmark events which have drastically shaped the trajectory of life and paved the way for the rich biodiversity we observe today (Smith and Szathmary 1997). Major evolutionary transitions include for instance the origin of eukaryotes from the merging of an archaeal host and a bacterial endosymbiont (McInerney et al. 2015; Zaremba-Niedzwiedzka et al. 2017; Donoghue et al. 2023) and the emergence of multicellularity (Ruiz-Trillo and Nedelcu 2015). Multicellularity has arisen several times independently in various eukaryotic lineages resulting in a diverse set of complex multicellular organisms, including brown algae, red algae, green algae and land plants, fungi, and animals (Ruiz-Trillo and Nedelcu 2015). The characteristics of the ancestral unicellular eukaryote and the mechanisms driving the emergence of multicellularity vary between lineages and remain subjects of ongoing research (Ruiz-Trillo and Nedelcu 2015).

The origin of animals through multicellularity has seen various hypotheses, each centred around the nature of the unicellular ancestor. This has been recently reviewed by Brunet and King (Brunet and King 2022) and is here summarized. Prior to the establishment of molecular phylogenies, proposed ancestral lineages spanned a range from amoebozoans (Haeckel 1876) to choanoflagellates (Metchnikoff 1886) and ciliates (Saville-Kent 1882). This lack of consensus throughout the 19th and 20th centuries, was amplified by both technical and conceptual limitations. A notable point of contention was the debate over animal monophyly. Some researchers questioned the relatedness of sponges to other animals, postulating the possibility of distinct ancestors for sponges (choanoflagellates) and the remainder of animals (ciliates) (Saville-Kent 1882). Contemporary molecular phylogenies unequivocally support the monophyly of animals and choanoflagellates as their sister group, together forming the clade Choanozoa, within the broader Holozoan clade (Wainright et al. 1993; Lang et al. 2002; Ruiz-Trillo et al. 2008). Choanozoa is corroborated by morphological and biochemical evidence: the collar complex surrounding the flagellum, a defining feature of choanoflagellates, is not only found in

sponge choanocytes but across various animals and is composed of cytoskeletal filaments that are homologous among choanoflagellates, sponges, and other animals (Nerrevang and Wingstrand 1970; Lyons 1973; Rieger 1976; Brunet and King 2017; Colgren and Nichols 2020). While the choanoflagellate-like ancestor hypotheses is now the most widely accepted, the specific mechanisms behind the evolution of animals from such an ancestor remain to be clarified. Theories have revolved around the two hypotheses of aggregative and clonal multicellularity, with the latter currently gaining wider acceptance (Brunet and King 2017). However, a recent theory posits that the mutual ancestor of animals and choanoflagellates presented a complex life-cycle, including transitions between amoeboid and flagellate phenotypes, similar to the cell types present in modern sponges (Arendt et al. 2015; Brunet and King 2017; Brunet et al. 2021; Brunet and King 2022). These alternative phenotypes were temporally segregated into different cells in the ancestor, however, following a process of clonal multicellularity these different phenotypes became spatially rather than temporally segregated. This combined with division of labour and innovation lead to the evolution of animals (Brunet and King 2017). The notion that living choanoflagellates present multiple phenotypes including sessile, swimming, and colonial forms, plus the fact that other closely related holozoans such as ichthyosporeans and filastereans also assume diverse cellular forms (Suga and Ruiz-Trillo 2013; Hehenberger et al. 2017; Parra-Acero et al. 2018; Brunet et al. 2019; Parra-Acero et al. 2020; Tikhonenkov, Hehenberger, et al. 2020; Tikhonenkov, Mikhailov, et al. 2020), support this line of investigation, that is currently topic of active research, driven by the emergence of holozoans as model organisms (Booth and King 2022).

Expansion of signal transduction systems in animals

Regardless of the precise mechanisms behind the origin of multicellularity in animals, this major transition has had profound implications. Obligate multicellular organisms such as animals must interact with the environment as a whole entity rather than as individual cells and this requires complex mechanisms for internal communication and coordination amongst cells. Consequently, cells must undergo subspecialisations for different tasks, whilst contemporarily maintaining the ability to collaborate with each other (Ruiz-Trillo et al. 2007). Ultimately this paved the way for the vast diversity of animal forms, ranging from relatively simple to extremely complex organisms with

intricate systems for self-coordination and interaction with the non-self, such as the nervous and immune systems (Bich et al. 2019; Jékely 2021; Jékely et al. 2021).

From a genetic perspective, we expect the emergence of novel genes to accompany the evolution of animals in response to these new challenges. Indeed, research into genes originating at the stem of metazoa point towards an increase in new genes for nucleic acid binding molecules, transcription factors and molecules involved in cell signalling (Paps and Holland 2018). Cell signalling plays a pivotal role in facilitating biological processes requiring communication amongst cells. Typically, it involves chemical messages or ligands—either endogenous or exogenous—that engage cellular receptors. This activation triggers a sequence of intracellular events, the signal transduction, involving second messengers and various effectors (Foreman et al. 2010). Ultimately, this allows cells to detect and react to extracellular cues either deriving from other cells, like hormones, neurotransmitters, and neuropeptides, or from external stimuli such as light (Elphick et al. 2018; Moroz et al. 2021; Oteiza and Baldwin 2021). There are many different types of receptors that generally fall within the categories of ligand-gated ion channels, enzyme linked receptors, G-protein-coupled receptors (GPCRs) and even intracellular receptors (Foreman et al. 2010). GPCRs, in particular, play a key role in numerous signalling pathways in animals, from neural communication, light reception and other sensory systems and immunity. Given the importance of cell signalling for animals, it is not surprising that one of the categories of gene families that was found to have significant emergence of new genes in the stem of metazoa is signalling molecules (Paps and Holland 2018). Additionally, even when there has not been a *de novo* origin of novel genes, there can be expansions within existing gene families effectively introducing novel genes that are often associated with new functions. This seems to have been the case for GPCR receptors. GPCR signalling is ancient, being present throughout eukaryotes, however, a huge expansion of this gene family occurs in animals. This is not seen neither in close relatives of animals nor in other multicellular organisms (e.g., plants have a comparatively limited set of GPCRs) (de Mendoza et al. 2014). This dramatic increase of GPCRs in animals is likely linked to their heightened need for rapid responsiveness to their environment.

Given the centrality of these receptors in orchestrating myriad biological processes, they have long been a primary subject of research, with a particular focus in deciphering their evolution to gain insights into the fundamental biological processes that they govern

(Fredriksson et al. 2003; Foster et al. 2019). Understanding the evolution of these molecules, sheds light on animal evolution, especially during its early stages when critical adaptations were likely to have occurred following the transition to the novel multicellular lifestyle. Similarly, unravelling the evolutionary histories of other molecules involved in GPCR signalling, such as the second messengers and effectors, is also important in understanding the evolution of cell signalling in animals.

General Aims of the Thesis

During my PhD, I was interested in investigating the evolution of signalling systems in animals. For this, I focused my attention on two different biological processes that rely on signal transduction systems. The first is vision, a widespread phenomenon in animals fundamental for the response to external light stimuli (Land and Nilsson 2012). The second is chemokine signalling, best known for its role in immunity but also involved in other physiological and developmental processes that require internal organismal communication (Murphy 2023). Each presented unique challenges but were both primarily addressed with phylogenetic methods and in some cases with additional bioinformatic approaches such as single cell sequencing analyses. In this short General Introduction, I will delineate the basic background and aims for both systems studied. In the next chapter, General Methods, I will introduce the basics of the methodologies used. Further details about both the background and the methodologies are then provided in the respective chapters.

The origin and evolution of vision in animals

Vision is an example of a sensory system that functions through GPCR signalling. It is a quintessential feature of animals, deeply influencing their ecology and behaviour (Nilsson 2009). At its core, vision consists of a photo-sensitive molecule coupled to a signal transduction machinery within a highly specialised photoreceptor cell. The photo-sensitive molecule is an opsin, a GPCR of class A, bound to a derivative of vitamin A, the retinal (Terakita 2005). When the retinal is hit by light it changes conformation (from 11-cis to all-trans), inducing a structural change of the opsin which in turn triggers the G alpha protein it is coupled with activating a signal transduction pathway called phototransduction. There are two major types of phototransduction, rhabdomeric and ciliary, depending on the type of opsins that initiate them, but both culminate in the modulation of ion channels initiating electrical signalling of the photoreceptor cell (Hardie and Juusola 2015; Lamb 2020).

Photoreceptor cells (PRCs) are classified based on the type of opsins and phototransduction pathway employed (Arendt 2003). A general peculiarity of PRCs is the

enlargement and folding of the membrane surface to increase the area with the photopigment and therefore enhance light sensitivity. This characteristic membrane folding is present within the cilia of ciliary PRCs of vertebrates, while in rhabdomeric PRCs of insects such as *Drosophila melanogaster* this folding is in the apical surface of the cell forming the rhabdomere (Arendt 2003). Historically, these morphological differences dictated PRC classification. It was believed that rhabdomeric PRCs were characteristic of the protostome (e.g. insects) lineage of Bilateria, while ciliary PRCs were specific to deuterostomes, including vertebrates (Eakin 1979). However, it is now known that ciliary PRCs are present within protostomes (Arendt et al. 2004; Passamaneck et al. 2011; von Döhren and Bartolomaeus 2018) and rhabdomeric PRCs within deuterostomes (Hattar et al. 2002; Ullrich-Lüter et al. 2011). Therefore, molecular definitions offer a more accurate classification, especially for non-bilaterian animals. While complex visual structures, such as eyes, are believed to have evolved independently on multiple occasions (Land and Nilsson 2012; Picciani et al. 2018), their fundamental units—photoreceptor cells—stem from a limited number of subtypes that may share a common ancestral cell type. This suggests that, despite variations in phototransduction machinery, there may exist a core set of regulatory genes defining this broad cell type (Arendt 2008; Arendt et al. 2016), consistent across all animal photoreceptor cells.

Beyond the phototransduction machinery and photoreceptor cells, vision encompasses another layer of molecular complexity. After the retinal is isomerized from it cis to its trans state by light, it must return to its cis state in order to be receptive to new light stimuli. This recycling occurs through a series of enzymatic reactions occurring as part of the retinol metabolism (Palczewski and Kiser 2020). As the opsin alone cannot carry out the visual function, this pathway that allows constant replenishment of the cis-retinal is just as essential part of the molecular assembly of vision.

Photoreceptor cells are present even in some early-branching animals, such as cnidarians (Nordström et al. 2003; Kozmik et al. 2008; Picciani et al. 2018) and potentially ctenophores (Horridge 1964; Jékely et al. 2015; Tamm 2016), suggesting that vision must have originated early in animal evolution. Some molecular components underpinning it, such as core signal transduction elements, likely trace back more anciently, while others, such as the regulatory genes involved in photoreceptor cell identity, may be animal innovations. Unravelling the evolutionary history of all these molecular players,

identifying key innovations and major family expansions, can not only elucidate the emergence of vision but also enrich our understanding of animal evolution more broadly.

Numerous studies have delved into the evolution of opsins, illuminating the vast diversity of these molecules across animals, including non-bilaterians (Feuda et al. 2012; Feuda et al. 2014; D'Aniello et al. 2015; Roberts et al. 2022; De Vivo et al. 2023; McCulloch et al. 2023). Such research has led to significant discoveries, including the identification of phylogenetically related placopsins in placozoans, a non-bilaterian phylum lacking neurons (Feuda et al. 2012). Nevertheless, comprehensive investigations into the evolution of all molecular components involved in vision remain sparse. Thus, one of my PhD goals was to fill in some of these gaps by investigating the evolution of the complex molecular assembly of vision. For this, I identified two main aims:

Aim 1: Reconstructing the evolution of the molecular components of photoreceptor cells.

The first aim is to understand the evolution of the molecular setup of photoreceptor cells, including both the phototransduction machinery and the regulatory toolkit that define the cell type. The objectives of aim 1 are addressed in Chapter 3.

Aim 2: Reconstructing the evolution of the retinol metabolism.

The second major aim is to investigate the evolution of the retinol metabolism that includes enzymes involved in the recovery of the cis-retinal, discerning whether specific components may have undergone distinct evolutionary events in animals. Aim 2 is addressed in Chapter 4.

Evolution and molecular diversity of chemokine signalling systems.

The immune system exemplifies an organism-wide system necessitating cellular coordination to detect and counteract external invaders. Present across the animal kingdom, immune systems function through an intricate range of subsystems (Yuan et al. 2014). In vertebrates, the chemokine signalling system plays a fundamental role in both innate and adaptive immunity (Murphy 2023). Best known for the chemoattraction of leukocytes during host defence (Wong and Fish 2003; Blanchet et al. 2012); chemokine signalling is also implicated in homeostasis, development (Zlotnik and Yoshie 2000; Tran and Miller 2003; López-Cotarelo et al. 2017), and neuronal communication (Tran and

Miller 2003; de Haas et al. 2007; Rostène et al. 2007). Failure of the system can lead to various diseases (Tran and Miller 2003; Blanchet et al. 2012), including cancer (Nagarsheth et al. 2017).

The chemokine system comprises two primary components: the chemokine ligands, small cytokines possessing chemotactic attributes, and the chemokine receptors, Class A GPCRs. Canonical chemokine ligands possess in their N-terminal portion characteristic cysteine patterns that can be used to classify them into subgroups. Canonical chemokine receptors are in turn classified based on the type of ligands they respond to, although there tends to be a high degree of promiscuity in the system (Zlotnik and Yoshie 2000; Nomiyama et al. 2011). Additionally, several other molecules have been implicated in the system. For example, ligands bearing varying degrees of sequence similarity to canonical chemokines have been found to activate some canonical receptors and/or have chemotactic properties (Zhang et al. 2018). Conversely, some so-called atypical chemokine receptors can bind canonical ligands, but do not trigger the signal transduction pathway necessary for chemokine function (Bonecchi and Graham 2016; Chen et al. 2018). Therefore, these additional players can be considered as "non-canonical" chemokine components. Yet, their relationship with the canonical components and amongst each other is unclear, hampering our understanding of the origin and evolution of the system. Applying evolutionary approaches can aid in clarifying the relatedness of these molecules and ultimately help clarify the evolution of the whole system. Thus, the second goal of my PhD was to explore the evolution of the chemokine system including both its canonical and non-canonical components. For this, three main aims were identified:

Aim 1: Uncovering the relationships among canonical and non-canonical components.

The first aim was to investigate the evolutionary relationships among all chemokine ligands and amongst all receptors, including all known canonical and non-canonical molecules. This served as a first step for subsequent phylogenetic analyses.

Aim 2: Reconstructing the evolution of all canonical and non-canonical ligands.

The following aim was to perform phylogenetic analyses for each ligand family identified to discern their evolution in animals.

Aim 3: Reconstructing the evolution of all canonical and non-canonical receptors.
Similarly, the last aim was to understand the evolutionary history of receptor groups.

The work addressing these aims is detailed in Chapter 5 and was carried out in collaboration with other members of the Feuda Group.

References

Arendt D. 2003. Evolution of eyes and photoreceptor cell types. *Int J Dev Biol* 47:563–571.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet* [Internet] 9:868–882. Available from: https://www.nature.com/articles/nrg2416

Arendt D, Benito-Gutierrez E, Brunet T, Marlow H. 2015. Gastric pouches and the mucociliary sole: setting the stage for nervous system evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 370:20150286. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2015.0286

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD, et al. 2016. The origin and evolution of cell types. *Nat Rev Genet* 17:744–757.

Arendt D, Tessmar-Raible K, Snyman H, Dorresteijn AW, Wittbrodt J. 2004. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science* 306:869–871.

Bich L, Pradeu T, Moreau J-F. 2019. Understanding Multicellularity: The Functional Organization of the Intercellular Space. *Frontiers in Physiology* [Internet] 10. Available from: https://www.frontiersin.org/articles/10.3389/fphys.2019.01170

Blanchet X, Langer M, Weber C, Koenen R, von Hundelshausen P. 2012. Touch of Chemokines. *Frontiers in Immunology* [Internet] 3:175. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2012.00175

Bonecchi R, Graham GJ. 2016. Atypical Chemokine Receptors and Their Roles in the Resolution of the Inflammatory Response. *Frontiers in Immunology* [Internet] 7:224. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2016.00224

Booth DS, King N. 2022. Chapter Three - The history of Salpingoeca rosetta as a model for reconstructing animal origins. In: Goldstein B, Srivastava M, editors. Current Topics in Developmental Biology. Vol. 147. Emerging Model Systems in Developmental Biology. Academic Press. p. 73–91. Available from: https://www.sciencedirect.com/science/article/pii/S0070215322000011

Brunet T, Albert M, Roman W, Coyle MC, Spitzer DC, King N. 2021. A flagellate-to-amoeboid switch in the closest living relatives of animals.Wittkopp PJ, Ruiz-Trillo I, López-García P, editors. *eLife* [Internet] 10:e61037. Available from: https://doi.org/10.7554/eLife.61037

Brunet T, King N. 2017. The Origin of Animal Multicellularity and Cell Differentiation. *Developmental Cell* [Internet] 43:124–140. Available from: https://www.cell.com/developmental-cell/abstract/S1534-5807(17)30769-4

Brunet T, King N. 2022. The Single-Celled Ancestors of Animals: A History of Hypotheses. In: The Evolution of Multicellularity. CRC Press.

Brunet T, Larson BT, Linden TA, Vermeij MJA, McDonald K, King N. 2019. Light-regulated collective contractility in a multicellular choanoflagellate. *Science* [Internet] 366:326–334. Available from: https://www.science.org/doi/full/10.1126/science.aay2346

Chen K, Bao Z, Tang P, Gong W, Yoshimura T, Wang JM. 2018. Chemokines in homeostasis and diseases. *Cell Mol Immunol* [Internet] 15:324–334. Available from: https://www.nature.com/articles/cmi2017134

Colgren J, Nichols SA. 2020. The significance of sponges for comparative studies of developmental evolution. *WIREs Developmental Biology* [Internet] 9:e359. Available from: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wdev.359

D'Aniello S, Delroisse J, Valero-Gracia A, Lowe EK, Byrne M, Cannon JT, Halanych KM, Elphick MR, Mallefet J, Kaul-Strehlow S, et al. 2015. Opsin evolution in the Ambulacraria. *Marine Genomics* [Internet] 24:177–183. Available from: https://www.sciencedirect.com/science/article/pii/S1874778715300349

De Vivo G, Crocetta F, Ferretti M, Feuda R, D'Aniello S. 2023. Duplication and Losses of Opsin Genes in Lophotrochozoan Evolution. *Molecular Biology and Evolution* [Internet] 40:msad066. Available from: https://doi.org/10.1093/molbev/msad066

von Döhren J, Bartolomaeus T. 2018. Unexpected ultrastructure of an eye in Spiralia: the larval ocelli of Procephalothrix oestrymnicus (Nemertea). *Zoomorphology* [Internet] 137:241–248. Available from: https://doi.org/10.1007/s00435-017-0394-3

Donoghue PCJ, Kay C, Spang A, Szöllősi G, Nenarokova A, Moody ERR, Pisani D, Williams TA. 2023. Defining eukaryotes to dissect eukaryogenesis. *Current Biology* [Internet] 33:R919–R929. Available from: https://www.sciencedirect.com/science/article/pii/S0960982223009879

Eakin RM. 1979. Evolutionary Significance of Photoreceptors: In Retrospect. *Am Zool* [Internet] 19:647–653. Available from: https://academic.oup.com/icb/article-lookup/doi/10.1093/icb/19.2.647

Elphick MR, Mirabeau O, Larhammar D. 2018. Evolution of neuropeptide signalling systems. *Journal of Experimental Biology* [Internet] 221:jeb151092. Available from: https://doi.org/10.1242/jeb.151092

Feuda R, Hamilton SC, McInerney JO, Pisani D. 2012. Metazoan opsin evolution reveals a simple route to animal vision. *PNAS* [Internet] 109:18868–18872. Available from: https://www.pnas.org/content/109/46/18868

Feuda R, Rota-Stabelli O, Oakley TH, Pisani D. 2014. The Comb Jelly Opsins and the Origins of Animal Phototransduction. *Genome Biology and Evolution* [Internet] 6:1964–1971. Available from: https://doi.org/10.1093/gbe/evu154

Foreman JC, Johansen T, Gibb AJ. 2010. Textbook of Receptor Pharmacology. CRC Press

Foster SR, Hauser AS, Vedel L, Strachan RT, Huang X-P, Gavin AC, Shah SD, Nayak AP, Haugaard-Kedström LM, Penn RB, et al. 2019. Discovery of Human Signaling Systems: Pairing Peptides to G Protein-Coupled Receptors. *Cell* [Internet] 179:895-908.e21. Available from: https://www.sciencedirect.com/science/article/pii/S0092867419311262

Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB. 2003. The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol Pharmacol* [Internet] 63:1256–1272. Available from: https://molpharm.aspetjournals.org/content/63/6/1256

de Haas AH, van Weering HRJ, de Jong EK, Boddeke HWGM, Biber KPH. 2007. Neuronal Chemokines: Versatile Messengers In Central Nervous System Cell Interaction. *Mol Neurobiol* [Internet] 36:137–151. Available from: https://doi.org/10.1007/s12035-007-0036-8

Haeckel E. 1876. The history of creation, or, The development of the earth and its inhabitants by the action of natural causes : doctrine of evolution in general, and of that of Darwin, Goethe, and Lamarck in particular / from the German of Ernst Haeckel ; the translation revised by E. Ray Lankester. London: Henry S. King Available from: https://www.biodiversitylibrary.org/item/99234

Hardie RC, Juusola M. 2015. Phototransduction in Drosophila. *Current Opinion in Neurobiology* [Internet] 34:37–45. Available from: https://www.sciencedirect.com/science/article/pii/S0959438815000173

Hattar S, Liao HW, Takao M, Berson DM, Yau KW. 2002. Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science* 295:1065–1070.

Hehenberger E, Tikhonenkov DV, Kolisko M, del Campo J, Esaulov AS, Mylnikov AP, Keeling PJ. 2017. Novel Predators Reshape Holozoan Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. *Current Biology* [Internet] 27:2043-2050.e6. Available from: https://www.sciencedirect.com/science/article/pii/S0960982217307078

Horridge GA. 1964. Presumed photoreceptive cilia in a ctenophore. *Quarterly Journal of microscopic science* [Internet]. Available from: https://openresearch-repository.anu.edu.au/handle/1885/167542

Jékely G. 2021. The chemical brain hypothesis for the origin of nervous systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 376:20190761. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0761

Jékely G, Godfrey-Smith P, Keijzer F. 2021. Reafference and the origin of the self in early nervous system evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 376:20190764. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0764

Jékely G, Paps J, Nielsen C. 2015. The phylogenetic position of ctenophores and the origin(s) of nervous systems. *EvoDevo* [Internet] 6:1. Available from: https://doi.org/10.1186/2041-9139-6-1

Kozmik Z, Ruzickova J, Jonasova K, Matsumoto Y, Vopalensky P, Kozmikova I, Strnad H, Kawamura S, Piatigorsky J, Paces V, et al. 2008. Assembly of the cnidarian camera-type eye from vertebrate-like components. *PNAS* [Internet] 105:8989–8993. Available from: https://www.pnas.org/content/105/26/8989

Lamb TD. 2020. Evolution of the genes mediating phototransduction in rod and cone photoreceptors. *Progress in Retinal and Eye Research* [Internet] 76:100823. Available from: https://www.sciencedirect.com/science/article/pii/S1350946219301107

Land MF, Nilsson D-E. 2012. Animal Eyes. Second Edition, Second Edition. Oxford, New York: Oxford University Press

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G. 2002. The Closest Unicellular Relatives of Animals. *Current Biology* [Internet] 12:1773–1778. Available from: https://www.sciencedirect.com/science/article/pii/S0960982202011879

López-Cotarelo P, Gómez-Moreira C, Criado-García O, Sánchez L, Rodríguez-Fernández JL. 2017. Beyond Chemoattraction: Multifunctionality of Chemokine Receptors in Leukocytes. *Trends in Immunology* [Internet] 38:927–941. Available from: https://www.sciencedirect.com/science/article/pii/S1471490617301655

Lyons KM. 1973. Collar cells in planula and adult tentacle ectoderm of the solitary coral Balanophyllia regia (anthozoa eupsammiidae). *Z.Zellforsch* [Internet] 145:57–74. Available from: https://doi.org/10.1007/BF00307189

McCulloch KJ, Babonis LS, Liu A, Daly CM, Martindale MQ, Koenig KM. 2023. Nematostella vectensis exemplifies the exceptional expansion and diversity of opsins in the eyeless Hexacorallia. *EvoDevo* [Internet] 14:14. Available from: https://doi.org/10.1186/s13227-023-00218-8

McInerney J, Pisani D, O'Connell MJ. 2015. The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 370:20140323. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0323

de Mendoza A, Sebé-Pedrós A, Ruiz-Trillo I. 2014. The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity. *Genome Biology and Evolution* [Internet] 6:606–619. Available from: https://doi.org/10.1093/gbe/evu038

Metchnikoff É. 1886. Embryologische Studien an Medusen : Ein Beitrag zur Genealogie der Primitiv-organe. Wien: A. Hölder Available from: https://www.biodiversitylibrary.org/item/27274

Moroz LL, Romanova DY, Kohn AB. 2021. Neural versus alternative integrative systems: molecular insights into origins of neurotransmitters. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 376:20190762. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0762

Murphy PM. 2023. 15 - Chemokines and Chemokine Receptors. In: Rich RR, Fleisher TA, Schroeder HW, Weyand CM, Corry DB, Puck JM, editors. Clinical

Immunology (Sixth Edition). New Delhi: Elsevier. p. 215–227. Available from: https://www.sciencedirect.com/science/article/pii/B9780702081651000150

Nagarsheth N, Wicha MS, Zou W. 2017. Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. *Nature Reviews Immunology* [Internet] 17:559–572. Available from: https://doi.org/10.1038/nri.2017.49

Nerrevang A, Wingstrand KG. 1970. On the Occurrence and Structure of Choanocyte-like Cells in Some Echinoderms. *Acta Zoologica* [Internet] 51:249–270. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1463-6395.1970.tb00436.x

Nilsson D-E. 2009. The evolution of eyes and visually guided behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 364:2833–2847. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2009.0083

Nomiyama H, Osada N, Yoshie O. 2011. A family tree of vertebrate chemokine receptors for a unified nomenclature. *Developmental & Comparative Immunology* [Internet] 35:705–715. Available from: https://www.sciencedirect.com/science/article/pii/S0145305X11000206

Nordström K, Wallén null, Seymour J, Nilsson D. 2003. A simple visual system without neurons in jellyfish larvae. *Proceedings of the Royal Society of London. Series B: Biological Sciences* [Internet] 270:2349–2354. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2003.2504

Oteiza P, Baldwin MW. 2021. Evolution of sensory systems. *Current Opinion in Neurobiology* [Internet] 71:52–59. Available from: https://www.sciencedirect.com/science/article/pii/S0959438821000969

Palczewski K, Kiser PD. 2020. Shedding new light on the generation of the visual chromophore. *Proc Natl Acad Sci U S A* [Internet] 117:19629–19638. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7443880/

Paps J, Holland PWH. 2018. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat Commun* [Internet] 9:1730. Available from: https://www.nature.com/articles/s41467-018-04136-5

Parra-Acero H, Harcet M, Sánchez-Pons N, Casacuberta E, Brown NH, Dudin O, Ruiz-Trillo I. 2020. Integrin-Mediated Adhesion in the Unicellular Holozoan Capsaspora owczarzaki. *Current Biology* [Internet] 30:4270-4275.e4. Available from: https://www.cell.com/current-biology/abstract/S0960-9822(20)31169-6

Parra-Acero H, Ros-Rocher N, Perez-Posada A, Kożyczkowska A, Sánchez-Pons N, Nakata A, Suga H, Najle SR, Ruiz-Trillo I. 2018. Transfection of Capsaspora owczarzaki, a close unicellular relative of animals. *Development* [Internet] 145:dev162107. Available from: https://doi.org/10.1242/dev.162107

Passamaneck YJ, Furchheim N, Hejnol A, Martindale MQ, Lüter C. 2011. Ciliary photoreceptors in the cerebral eyes of a protostome larva. *EvoDevo* [Internet] 2:6. Available from: https://doi.org/10.1186/2041-9139-2-6

Picciani N, Kerlin JR, Sierra N, Swafford AJM, Ramirez MD, Roberts NG, Cannon JT, Daly M, Oakley TH. 2018. Prolific Origination of Eyes in Cnidaria with Co-

option of Non-visual Opsins. *Current Biology* [Internet] 28:2413-2419.e4. Available from: https://www.sciencedirect.com/science/article/pii/S0960982218306912

Rieger RM. 1976. Monociliated epidermal cells in Gastrotricha: Significance for concepts of early metazoan evolution. *Journal of Zoological Systematics and Evolutionary Research* [Internet] 14:198–226. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0469.1976.tb00937.x

Roberts NS, Hagen JFD, Johnston RJ. 2022. The diversity of invertebrate visual opsins spanning Protostomia, Deuterostomia, and Cnidaria. *Developmental Biology* [Internet] 492:187–199. Available from: https://www.sciencedirect.com/science/article/pii/S0012160622002007

Rostène W, Kitabgi P, Parsadaniantz SM. 2007. Chemokines: a new class of neuromodulator? *Nat Rev Neurosci* [Internet] 8:895–903. Available from: https://www.nature.com/articles/nrn2255

Ruiz-Trillo I, Burger G, Holland PWH, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. *Trends in Genetics* [Internet] 23:113–118. Available from: https://www.cell.com/trends/genetics/abstract/S0168-9525(07)00023-6

Ruiz-Trillo I, Nedelcu AM. 2015. Evolutionary Transitions to Multicellular Life: Principles and Mechanisms edited by Iñaki Ruiz-Trillo and Aurora M. Nedelcu. *Advances in Marine Genomics 2. Springer* [Internet] 91:370–371. Available from: https://www.journals.uchicago.edu/doi/abs/10.1086/688137

Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A Phylogenomic Investigation into the Origin of Metazoa. *Molecular Biology and Evolution* [Internet] 25:664–672. Available from: https://doi.org/10.1093/molbev/msn006

Saville-Kent W. 1882. A Manual of the Infusoria: Including a Description of All Known Flagellate, Ciliate, and Tentaculiferous Protozoa, British and Foreign, and an Account of the Organization and the Affinities of the Sponges. D. Bogue

Smith T late PJM, Szathmary E. 1997. The Major Transitions in Evolution. Oxford, New York: Oxford University Press

Suga H, Ruiz-Trillo I. 2013. Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. *Developmental Biology* [Internet] 377:284–292. Available from: https://www.sciencedirect.com/science/article/pii/S0012160613000146

Tamm SL. 2016. Novel Structures Associated with Presumed Photoreceptors in the Aboral Sense Organ of Ctenophores. *Biol Bull* 231:97–102.

Terakita A. 2005. The opsins. *Genome Biology* [Internet] 6:213. Available from: https://doi.org/10.1186/gb-2005-6-3-213

Tikhonenkov DV, Hehenberger E, Esaulov AS, Belyakova OI, Mazei YA, Mylnikov AP, Keeling PJ. 2020. Insights into the origin of metazoan multicellularity from predatory unicellular relatives of animals. *BMC Biology* [Internet] 18:39. Available from: https://doi.org/10.1186/s12915-020-0762-1

Tikhonenkov DV, Mikhailov KV, Hehenberger E, Karpov SA, Prokina KI, Esaulov AS, Belyakova OI, Mazei YA, Mylnikov AP, Aleoshin VV, et al. 2020. New Lineage of Microbial Predators Adds Complexity to Reconstructing the Evolutionary Origin of Animals. *Current Biology* [Internet] 30:4500-4509.e5. Available from: https://www.cell.com/current-biology/abstract/S0960-9822(20)31251-3

Tran PB, Miller RJ. 2003. Chemokine receptors: signposts to brain development and disease. *Nature Reviews Neuroscience* [Internet] 4:444–455. Available from: https://doi.org/10.1038/nrn1116

Ullrich-Lüter EM, Dupont S, Arboleda E, Hausen H, Arnone MI. 2011. Unique system of photoreceptors in sea urchin tube feet. *Proc Natl Acad Sci U S A* 108:8367–8372.

Wainright PO, Hinkle G, Sogin ML, Stickel SK. 1993. Monophyletic Origins of the Metazoa: an Evolutionary Link with Fungi. *Science* [Internet] 260:340–342. Available from: https://www.science.org/doi/10.1126/science.8469985

Wong MM, Fish EN. 2003. Chemokines: attractive mediators of the immune response. *Seminars in Immunology* [Internet] 15:5–14. Available from: https://www.sciencedirect.com/science/article/pii/S1044532302001239

Yuan S, Tao X, Huang S, Chen S, Xu A. 2014. Comparative Immune Systems in Animals. *Annual Review of Animal Biosciences* [Internet] 2:235–258. Available from: https://doi.org/10.1146/annurev-animal-031412-103634

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* [Internet] 541:353–358. Available from: https://www.nature.com/articles/nature21031

Zhang K, Shi S, Han W. 2018. Research progress in cytokines with chemokine-like function. *Cellular & Molecular Immunology* [Internet] 15:660–662. Available from: https://doi.org/10.1038/cmi.2017.121

Zlotnik A, Yoshie O. 2000. Chemokines: A New Classification System and Their Role in Immunity. *Immunity* [Internet] 12:121–127. Available from: https://doi.org/10.1016/S1074-7613(00)80165-X

Chapter 2

General Methods

General Methods

To address the broad research questions of my thesis – the evolution of vision and the evolution of chemokine signalling – I used various bioinformatic methodologies. While detailed methods are described in each respective chapter, several basic approaches were shared amongst the different projects. Phylogenetic methods were applied in all projects, and one project additionally incorporated some analyses of single-cell sequencing data. In this chapter I will provide a basic overview of the methodologies, which will serve as a common foundation for the next chapters.

Phylogenetic analyses

All aims within this thesis required phylogenetic analysis of gene families essential to the biological processes of interest. The main steps common to Chapters 3, 4, and 5 are outlined here.

Dataset preparation

Obtaining starting queries

The first elementary step involves determining which gene families to explore with phylogenetic studies and to obtain reliable reference sequences to use as starting queries for the analyses. While literature serves as a foundational reference, leveraging pathway databases can ensure comprehensive coverage of essential components, especially when examining extensive pathways. One such pathway database is KEGG, which also provides lists of known homologs for pathway components (Kanehisa 2019; Kanehisa et al. 2021). I utilized KEGG as an initial source for reference sequences in Chapters 3 (evolution of phototransduction and photoreceptor cells) and 4 (evolution of retinol metabolism). For Chapter 5 (evolution of chemokine signalling), the primary database of reference was Guide to Pharmacology Database (Bachelerie et al. 2020). For all projects, a supplementary source for reference sequences was UniProt (Boutet et al. 2016; Poux et al. 2017; The UniProt Consortium 2023).

Choice of species

The comparative analysis of systems and signalling pathways requires the examination of genomes and predicted proteomes across a diverse spectrum of species. Thus, an essential preliminary step is selecting the species that best fit the research context. A primary consideration is determining the appropriate taxonomic sampling based on the research question. For example, in Chapters 3 and 4, primary focus was on early branching animals and closest relatives of animals, reflecting the onset of vision in the early stages of animal evolution. Yet, given the possibly ancient origin of certain components of the pathways under study, it was crucial to incorporate representatives from all major eukaryotic lineages. For this, my primary references were Adl et al. 2019 for eukaryotic classification and Burki et al. 2020 for phylogenetic relationships (Adl et al. 2019; Burki et al. 2020). In contrast, the chemokine signalling system is known only in vertebrates, with some non-canonical components potentially existing in other bilaterians. As such, in Chapter 5, species sampling was limited to animals, with an emphasis on vertebrates, a balanced representation of other bilaterians and a few non-bilaterians for a comprehensive search. Another vital consideration in species selection is the quality of available genomes/proteomes. The quality of the predicted proteome can significantly impact the outcomes and reliability of subsequent bioinformatic analyses. High-quality genomes, which are characterised by high levels of completeness and accuracy offer a more reliable representation of an organism's genetic blueprint. Errors, contamination or ambiguities in the sequence can lead to false or missed identifications, impacting downstream analyses (Simion et al. 2018; Waterhouse et al. 2018; Manni et al. 2021; Simakov et al. 2022). A hallmark of high-quality proteomes is their completeness. If a gene family is not identified in a species with a high-quality complete proteome, then it likely reflects true absence and not a technical limitation. In certain scenarios, there might be key species essential to the study, that may have a proteome with low level of completeness. To compensate for this, the solution is to incorporate multiple closely related species, thereby amplifying the chances of detecting the presence of specific gene families within that taxonomic lineage. The tool I used to assess the proteome completeness was BUSCO (Benchmarking Universal Single-Copy Orthologs) (Waterhouse et al. 2018; Manni et al. 2021). BUSCO searches the proteomes for a list of genes that are known to be universally present in single copy (the "BUSCO" genes) within a taxon. It scans the dataset using lineage-specific BUSCO profiles built using hidden Markov models (HMMs), statistical models that can capture the patterns in a set

of sequences (Krogh et al. 1994). The choice of lineage-specific BUSCO profiles for the search depends on the organisms under study. For example, in Chapters 3 and 4 I employed the BUSCO profiles designed for eukaryotes, whereas in Chapter 5, I utilized those tailored for metazoans. By providing the percentage of complete BUSCOs identified in each proteome searched, it offers a quantitative measure of the completeness of a dataset in terms of expected gene content. It also differentiates between complete BUSCO genes found in single versus multiple copies. As BUSCO genes are expected to be found in single copy, a high percentage of multi-copy complete BUSCOs may be an indicator of assembly issues. It also assesses the percentage of fragmented and missing BUSCOs, thereby providing a full picture of the proteome completeness. Combining this rigorous assessment with taxonomic considerations, it was possible to build tailored species databases for each Chapter, ensuring the robustness of the subsequent analyses.

Phylogenetic analyses

Initial sequence similarity-based data mining

The collected queries can be used to identify within the species database, homologous sequences to be used for the phylogenetic analyses. This "data mining" step can be first approached through sequence similarity methods. For this, I used BLAST (Basic Local Alignment Search Tool) for amino acid sequences (Altschul et al. 1997; Camacho et al. 2009). This widely used tool works by searching for an initial short match between the query and the database sequence, after which it attempts to add adjacent amino acids to extend the hit. As the alignment grows it is scored based on the exactness of the match, the extension stops if the score drops below a certain fraction of the highest score. BLAST retains this local alignment if its highest score has an expected value (e-value) below a user defined threshold (Lemey et al. 2009). The resulting hits are therefore considered to be more similar to each other than would be expected by chance, suggesting probable homology. This is a very powerful tool to narrow down potential homologs from large protein databases. The choice of e-value cut-off is critical, as if it is too loose (high) unrelated sequences may be collected, while if it is to strict (low) potential homologs might be missed. The e-value is influenced by the query length and database size: shorter queries and larger databases increase the probability of random hits; therefore, the e-value will tend to be higher in these cases. Given these complexities and recognizing that an optimal e-value might differ across gene families, in this thesis I adopted a strategy of

initiating with relatively loose BLAST searches followed by additional methodologies to further refine the results.

Optimisation of final gene family datasets

While BLAST served as the foundational method in all my chapters and is in general a very common tool, additional refinement of the gene families can be obtained by diverse strategies. In this thesis, the strategies employed can roughly fit into two categories: targeted versus large-scale approaches. In the first instance, *ad hoc* information about each gene family of interest is used to refine the search. This strategy was employed in Chapter 3, where I refined my BLAST results by a combination of two targeted approaches. Initially, I re-ran BLAST against SwissProt (Boutet et al. 2016; Poux et al. 2017), a high-quality curated database of annotated sequences, retaining only those sequences that correctly matched the desired gene family within the top hits. Subsequently, I filtered sequences by identifying known protein domains typical of each protein family. Further details can be found in the Methods section of Chapter 3. This is a highly precise strategy ensuring high confidence results; however, it is time consuming and requires a thorough knowledge of the gene families. The alternative approach employs sequence clustering tools to discern the relatedness among sequences, which is advantageous for broader albeit less targeted comparisons. This approach helps filter out unrelated sequences that were initially identified by BLAST but appear unrelated with the rest of the cluster. It also aids in distinguishing sub-families within a larger superfamily and clarifying connections amongst families previously classified solely by function rather than by evolutionary relationships. Different methods employ this clustering strategy. In Chapter 5, I utilized CLANS (Frickey and Lupas 2004), a tool that simply clusters sequences based on all-vs-all BLAST scores. Conversely, Chapter 4 employed more sophisticated methods that combine various clustering, phylogenetic and network analyses algorithms to infer orthogroups of sequences (further details are available in Chapter 4).

Annotating Sequences

A useful additional step is to provide annotations to the sequences collected as not all species proteomes are annotated to start with. To efficiently navigate large trees or sequence clusters and annotate their clades and groups, it is advantageous to have as many sequences as possible already with a “name”. Even for sequences from model organisms that come pre-annotated, nomenclature can vary greatly among species, complicating the

rapid identification of a clade or cluster. To address this, it is useful to standardize sequence naming. In this thesis, a common approach to achieve this was by BLASTing all sequences against SwissProt and retaining the top hit as the annotation. While this is not always precise, it provides a quick preliminary naming system. In some cases, more detailed annotation decisions might require manual inspection of sequences. Taxon-specific databases can be useful for this. Throughout this thesis, frequently consulted databases included: GeneCards for *Homo sapiens* (Stelzer et al. 2016); MGI for *Mus musculus* (Blake et al. 2021); FlyBase for *Drosophila melanogaster* (Larkin et al. 2021); Echinobase for *Strongylocentrotus purpuratus* and other echinoderms (Arshinoff et al. 2022); TAIR for *Arabidopsis thaliana* (Berardini et al. 2015).

Multiple sequence alignment and trimming

After the optimal curation of final gene families, the subsequent step involves aligning the sequences. This is a fundamental step in phylogenetic analyses. The underlying principle is that if sequences are homologous, each amino acid position traces back to a shared ancestral state and sequences can be aligned in such a way that each column represents homologous positions. In the resulting alignment, some positions might be highly conserved, while others divergent. Additionally, due to deletion or insertion events, homologous sequences can vary in length, leading to gaps for some sequences in the alignment. Overall, the alignment captures the evolutionary changes the sequences have undergone (Lemey et al. 2009). Multiple sequence alignments throughout this thesis were constructed using the MAFFT software (Katoh et al. 2002; Katoh and Standley 2013). The reliability and accuracy of multiple sequence alignments are critical for the quality of subsequent phylogenetic analyses. Removing poorly aligned regions from an alignment can enhance the quality of these analyses. Throughout this thesis the trimAl software has been used to trim alignments based on gap cut-offs and automatically computed parameters (Capella-Gutiérrez et al. 2009).

Inferring phylogenetic trees for each gene family

The multiple sequence alignment serves as foundation for constructing the phylogenetic tree for the gene family under examination. The method used to construct phylogenetic trees throughout this thesis is maximum likelihood using the software IQTREE2 (Hoang et al. 2018; Minh et al. 2020). This method aims to find the tree topology that best explains the observed data (i.e., the sequence alignment) given a particular model of sequence evolution. For a given tree and model, the likelihood is the probability of observing the

sequence alignment, given that tree. Maximum likelihood algorithms search the space of possible tree topologies to find the one that has the highest likelihood. The tree with the highest likelihood is considered the best estimate of the true phylogeny (Felsenstein 2003; Lemey et al. 2009). Models of protein evolution describe patterns and rates of amino acid substitutions and are used to estimate evolutionary distances between sequences. Although all models factor in attributes like the biochemical properties of amino acids, they can diverge in their utilization of specific substitution matrices and other parameters, such as rate variations across sites and differences in amino acid frequencies. Such distinctions make certain models more apt for specific datasets or evolutionary contexts (Felsenstein 2003; Lemey et al. 2009). To ensure the optimal model selection for each gene family in this thesis, I utilized the model finder feature of ITREE2 (Kalyaanamoorthy et al. 2017). To assess the confidence of the relationships recovered through phylogenetic tree inference, it is useful to calculate branch supports. Throughout my thesis I mainly used the IQTREE2 ultrafast bootstrap approximation method (Minh et al. 2013; Hoang et al. 2018) with 1000 replicates. This method is a computationally efficient alternative to the traditional bootstrap (Felsenstein 1985; Felsenstein 2003). While the conventional approach resamples the alignment dataset to produce pseudo-replicate datasets, infers respective trees and gauges support for branches based on the frequency of their appearance, the ultrafast bootstrap method streamlines this by approximating the process without fully resampling the dataset for each replicate. Additionally, in Chapter 5, to address the challenges of constructing trees for short, rapidly evolving sequences such as chemokines, the transfer bootstrap expectation (TBE) method (Lemoine et al. 2018) was also used. TBE assesses branch support by allowing for slight variations in the placement of sequences within the bootstrap trees, focusing more on the preservation of the main groupings or splits. If these primary relationships are consistent, the branch receives support, even if there are minor differences.

Species trees

In addition to the gene trees, some subsequent analyses, such as gene tree-species tree reconciliations (see below), also require species trees. The species trees constructed in this thesis are not intended to resolve phylogenetic relationships among the species studied. Instead, the primary goal was to have a species tree comprising the specific set of species used for the gene trees, serving as a reference where species relationships information was needed. To construct these species trees, I leveraged BUSCO results.

BUSCO identifies the complete single-copy BUSCOs in each analysed species and provides the sequences for these genes in each species. These BUSCO genes can be used to create a supermatrix for the species tree. The tree-building followed a maximum likelihood approach, after identifying the best-fit model as described above.

Gene tree to species tree reconciliation

In some cases, it is useful to re-estimate gene trees in light of known species relationships, as the histories of gene trees are intrinsically linked to the species tree. Gene tree to species tree reconciliation methods, which account for this relationship, can enhance tree inference, especially when phylogenetic signal is weak (Boussau and Scornavacca 2020; Williams et al. 2023). In this thesis, the GeneRax software (Morel et al. 2020) was used to reconcile gene trees to species trees. GeneRax re-infers the gene tree using maximum likelihood, guided by the species tree. Additionally, this reconciliation elucidates speciation, duplication, and loss events at each node of the gene tree. Such insights are invaluable for distinguishing between paralogs (genes that originate from a duplication event) and orthologs (genes that originate from a speciation event). Furthermore, thanks to the information about species relationships, it is also possible to accurately root gene trees, a challenge that is often complex without such context.

Analyses of single-cell sequencing data

For one of my aims— understanding the molecular setup of photoreceptor cells (Chapter 3)—I also I incorporated single-cell sequencing analyses of publicly available data. Specifically, after having determined the presence or absence of phototransduction genes in the genomes of target species, the next objective was to determine if these genes were co-expressed within a single cell type, that could represent a photoreceptor cell. Additionally, the aim was to uncover shared genetic patterns prevalent in animal photoreceptor cells, with an emphasis on regulatory genes. Single-cell RNA sequencing is a technique that is used to profile gene expression at the level of individual cells, therefore, analysing publicly available data for various animals has the potential to answer these questions. In Chapter 3 I combined the use of single-cell analyses software and some *ad hoc* strategies designed for the specific research question. While the precise methodologies are detailed in the Methods of Chapter 3, here I will provide a brief overview of the principles guiding the main steps.

Preliminary steps

Choice of species and obtaining datasets

The choice of species was guided by similar considerations as for the phylogenetic analyses: since vision via photoreceptor cells likely emerged during the early history of animals, the ideal dataset would include a balanced representation of major animal clades with emphasis on non-bilaterians. In practice though, the selection of species for analysis was primarily driven by the availability of published single-cell data. Although single-cell sequencing is gaining traction and new datasets spanning tissues, organs, and entire organisms are consistently emerging, the volume of such data is still currently quite limited, especially for non-model organisms. At the time of starting the work for Chapter 3, I was able to identify 12 species for the single-cell analysis, including 7 species spanning all four non-bilaterian phyla. The authors of the publications for all these species had already performed the preliminary steps to process the results from their sequencing: therefore, reads were already mapped to reference genomes and gene to cells count matrices computed. For all the species datasets, I downloaded the molecular count matrices, that was the input needed for the subsequent clustering step (see below).

Clustering cells into "metacells"

A typical step in single-cell sequencing analyses is to group cells into clusters based on similar expression profiles. The appropriate method for this clustering often depends on the dataset's specifics and the research question at hand. However, a common challenge in this step is addressing the intrinsic variability and noise present in single-cell data (Baran et al. 2019). One major source of technical noise is introduced through partial sampling of the RNA within a cell. This technical variance obscures the true biological variance. This issue becomes particularly problematic in datasets with low sequencing coverage, such as those from whole organisms that encompass numerous cell types. One method, MetaCell (Baran et al. 2019), addresses this limitation by inferring "metacells". A metacell is defined as a group of single-cell sequencing profiles that, statistically, could be seen as deriving from the RNA pool from a single cell. It is therefore a representation of a cell state. These metacells then act as foundational units for portraying complex gene expression patterns and for modelling subtle molecular states. In Chapter 3, I followed the default MetaCell R pipeline provided by the authors. The core steps are the identification of feature genes based on gene distributions statistics; the construction of a

similarity k-nn graph to connect pairs of cells on the basis of the feature genes; a resampling of the graph to obtain a co-clustering graph based on how often pairs of cells co-occurred. Further refinement is obtained by filtering outliers and splitting metacells with strong sub-cluster structure.

Identifying photoreceptor cells and cross species comparisons

Identification of candidates PRCs

Once metacells are computed, the next objective is to identify if some of them and which ones may present a photoreceptor (PRC)-like profile. The strategy I used relied on identifying metacells with high opsin expression combined with the expression of other phototransduction genes as additional markers. Further details are in the Methods of Chapter 3.

Exploration of the regulatory genes expressed in candidate PRCs

The subsequent step involved extracting all the genes expressed in each candidate PRC of all species and identifying "regulatory genes", including, for instance, transcription factors that are important for determining cell type identity. A comprehensive explanation of this procedure is provided in Chapter 3.

Comparisons across species

The final stage of my analysis consisted in performing all-against-all comparisons of all PRC metacells from all species to uncover patterns of shared regulatory genes expression. This analysis was performed at various levels of confidence by comparing both the shared genes that are most highly expressed in metacells and genes that are expressed but at lower expression levels. To gain deeper insights into the categories of regulatory genes consistently conserved across diverse species, I quantified the proportions of transcription factors, cofactors, and other regulatory genes present. Additionally, I identified which transcription factor families and DNA-binding domains were most prevalent in the dataset. A comprehensive breakdown of this process is detailed in the Methods section of Chapter 3.

References

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al. 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology* [Internet] 66:4–119. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jeu.12691

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* [Internet] 25:3389–3402. Available from: https://doi.org/10.1093/nar/25.17.3389

Arshinoff BI, Cary GA, Karimi K, Foley S, Agalakov S, Delgado F, Lotay VS, Ku CJ, Pells TJ, Beatman TR, et al. 2022. Echinobase: leveraging an extant model organism database to build a knowledgebase supporting research on the genomics and biology of echinoderms. *Nucleic Acids Research* [Internet] 50:D970–D979. Available from: https://doi.org/10.1093/nar/gkab1005

Bachelerie F, Ben-Baruch A, Burkhardt AM, Charo IF, Combadiere C, Förster R, Farber JM, Graham GJ, Hills R, Horuk R, et al. 2020. Chemokine receptors (version 2020.5) in the IUPHAR/BPS Guide to Pharmacology Database. *IUPHAR/BPS Guide to Pharmacology CITE* [Internet] 2020. Available from: http://journals.ed.ac.uk/gtopdb-cite/article/view/5178

Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, Meir Z, Hoichman M, Lifshitz A, Tanay A. 2019. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology* [Internet] 20:206. Available from: https://doi.org/10.1186/s13059-019-1812-2

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *genesis* [Internet] 53:474–485. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/dvg.22877

Blake JA, Baldarelli R, Kadin JA, Richardson JE, Smith CL, Bult CJ, the Mouse Genome Database Group. 2021. Mouse Genome Database (MGD): Knowledgebase for mouse–human comparative biology. *Nucleic Acids Research* [Internet] 49:D981–D987. Available from: https://doi.org/10.1093/nar/gkaa1083

Boussau B, Scornavacca C. 2020. Reconciling Gene trees with Species Trees. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.2:1-3.2:23. Available from: https://hal.science/hal-02535529

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. Methods in Molecular Biology. New York, NY: Springer. p. 23–54. Available from: https://doi.org/10.1007/978-1-4939-3167-5_2

Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The New Tree of Eukaryotes. *Trends in Ecology & Evolution* [Internet] 35:43–55. Available from: https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(19)30257-5

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* [Internet] 10:421. Available from: https://doi.org/10.1186/1471-2105-10-421

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* [Internet] 25:1972–1973. Available from: https://doi.org/10.1093/bioinformatics/btp348

Felsenstein J. 1985. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution* [Internet] 39:783–791. Available from: https://doi.org/10.1111/j.1558-5646.1985.tb00420.x

Felsenstein J. 2003. Inferring Phylogenies. Oxford, New York: Oxford University Press

Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* [Internet] 20:3702–3704. Available from: https://doi.org/10.1093/bioinformatics/bth444

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* [Internet] 35:518–522. Available from: https://doi.org/10.1093/molbev/msx281

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* [Internet] 14:587–589. Available from: https://www.nature.com/articles/nmeth.4285

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Science* [Internet] 28:1947–1951. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3715

Kanehisa M, Sato Y, Kawashima M. 2021. KEGG mapping tools for uncovering hidden features in biological data. *Protein Science* [Internet] n/a. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4172

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* [Internet] 30:3059–3066. Available from: https://doi.org/10.1093/nar/gkf436

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* [Internet] 30:772–780. Available from: https://doi.org/10.1093/molbev/mst010

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology* [Internet] 235:1501–1531. Available from: https://www.sciencedirect.com/science/article/pii/S0022283684711041

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, et al. 2021. FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research* [Internet] 49:D899–D907. Available from: https://doi.org/10.1093/nar/gkaa1026

Lemey P, Salemi M, Vandamme A-M eds. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd ed. Cambridge: Cambridge University Press Available from: https://www.cambridge.org/core/books/phylogenetic-handbook/A9D63A454E76A5EBCCF1119B3C56D766

Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* [Internet] 556:452–456. Available from: https://www.nature.com/articles/s41586-018-0043-0

Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols* [Internet] 1:e323. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.323

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution* [Internet] 30:1188–1195. Available from: https://doi.org/10.1093/molbev/mst024

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* [Internet] 37:1530–1534. Available from: https://doi.org/10.1093/molbev/msaa015

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution* [Internet] 37:2763–2774. Available from: https://doi.org/10.1093/molbev/msaa141

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* [Internet] 33:3454–3460. Available from: https://doi.org/10.1093/bioinformatics/btx439

Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, O'Connell BL, Dear P, Martinez DE, Steele RE, et al. 2022. Deeply conserved synteny and the evolution of metazoan chromosomes. *Science Advances* [Internet] 8:eabi5884. Available from: https://www.science.org/doi/10.1126/sciadv.abi5884

Simion P, Belkhir K, François C, Veyssier J, Rink JC, Manuel M, Philippe H, Telford MJ. 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology* [Internet] 16:28. Available from: https://doi.org/10.1186/s12915-018-0486-7

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, et al. 2016. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*

[Internet] 54:1.30.1-1.30.33. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.5

The UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* [Internet] 51:D523–D531. Available from: https://doi.org/10.1093/nar/gkac1052

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* 35:543–548.

Williams TA, Davin AA, Morel B, Szánthó LL, Spang A, Stamatakis A, Hugenholtz P, Szöllősi GJ. 2023. The power and limitations of species tree-aware phylogenetics. :2023.03.17.533068. Available from: https://www.biorxiv.org/content/10.1101/2023.03.17.533068v1

Chapter 3

The Molecular Evolution of Animal Phototransduction and Photoreceptor Cells

Abstract

The origin of vision has been a major novelty in animals, playing a fundamental role in the evolution of complex behaviours, such as mate choice and predator avoidance, that distinguish animals from other organisms. Vision starts with a light-triggered phototransduction cascade that occurs in specialised neurons, known as the photoreceptor cells (PRCs). The two main PRC types, ciliary and rhabdomeric, employ specific as well as common genes for phototransduction. While fundamental for vision, the origin and evolution of photoreceptor cells and their phototransduction pathways are still unclear.

Using phylogenetic methods, including gene-tree to species-tree reconciliations, I studied the pattern of gene duplications for all phototransduction genes in more than 80 species, including non-bilaterian metazoans and other eukaryotes. Next, I investigated the expression of phototransduction genes in available single-cell RNA-sequencing data of various animals, including non-bilaterians. Using phototransduction genes as markers, putative photoreceptor-like cells were identified across animals and their regulatory toolkits were compared.

Phototransduction gene families were found to be generally very ancient, predating the origin of vision. Major family expansions and diversifications often occurred just prior to or at the base of animals, reflecting the ability of animals to broaden their responses to the environment. Moreover, putative photoreceptor cells identified in non-bilaterians appeared to express some but not all components of the two well characterised phototransduction pathways, suggesting potential lineage-specific components involved in phototransduction. Finally, most of the regulatory genes shared across animal PRCs are transcription factors, with the most predominant families including bZIP transcription factor, zinc finger C2H2 and homeobox. While several regulatory genes are recurrent throughout animals, the exact same combinations of these genes rarely span all phyla.

Introduction

Animal evolution has gone hand in hand with an increasing refinement of the ability to sense and respond to the environment. One fundamental sense is vision, which is widespread throughout the animal kingdom (Nilsson 2009; Nilsson 2013). At a molecular level, the visual process begins with the reception of light by a photosensitive molecule. This light-activated molecule in turn triggers a chain of molecular signalling within the cell that culminates into ion channel opening/closing resulting in electrical signalling. This phototransduction process occurs within specialised neurons called photoreceptor cells (Nilsson 2009).

The photosensitive molecule is composed of an opsin, a membrane bound G-protein coupled receptor (GPCR), and a light-sensitive chromophore bound to it (Terakita 2005). This chromophore, the retinal, derives from the metabolism of vitamin A. In the dark, the retinal is in its 11-cis state. When hit by light photons, it isomerizes into its all-trans state, inducing the structural change in the opsin that in turn initiates the phototransduction pathway (Terakita 2005; Palczewski and Kiser 2020; Widjaja-Adhi and Golczak 2020).

Two alternative phototransduction cascades have been described in detail (Yau and Hardie 2009). In *Drosophila melanogaster* (Figure 3.1A), the opsin activates a Gq-type G protein. The alpha subunit detaches from the complex and activates phospholipase C beta that initiates a phosphoinositide cascade. This results in the opening of transient receptor potential (trp) and trp-like (trpl) channels with consequent depolarization of the cell (Wang and Montell 2007; Hardie and Juusola 2015). Whereas in vertebrates, as exemplified by *Homo sapiens* (Figure 3.1B), the opsin activates transducin (Gt) a G protein of the Gi/o-type that activates phosphodiesterase 6 (PDE6) that hydrolyses cyclic GMP. The drop in cGMP levels causes the cyclic nucleotide gated ion channels (CNGCs) to close, followed by a hyperpolarization of the cell (Lamb 2020). Some molecular components are shared between both pathways, whilst others are specific to either one or the other pathway (Figure 3.1 and Table 3.1). Reconstructing the evolutionary history of each phototransduction gene family is necessary to understand when the complete phototransduction pathways originated and may have started to acquire their visual function.

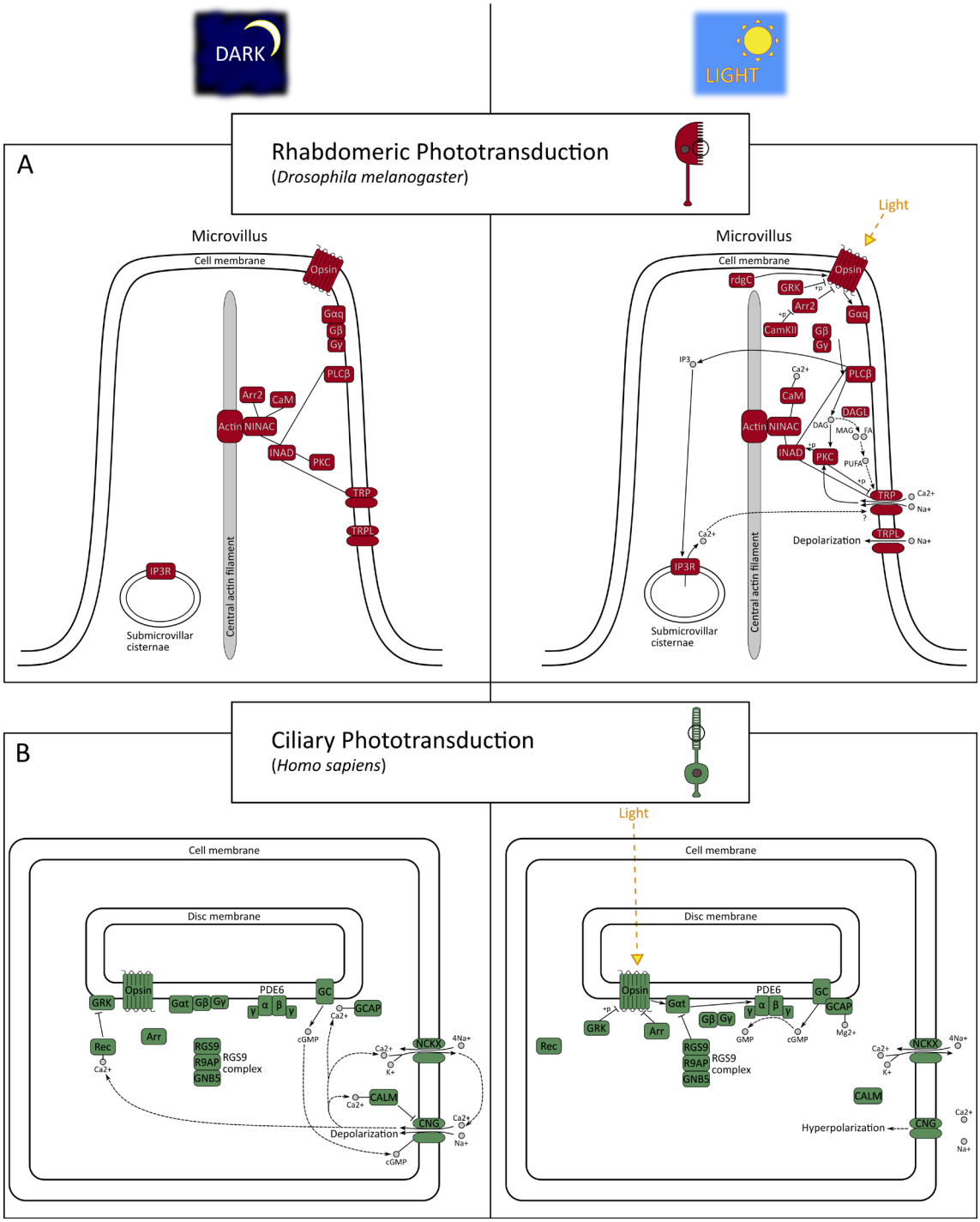


Figure 3.1. Schematics of rhabdomeric and ciliary phototransduction pathways. (**A**) Rhabdomeric phototransduction in *Drosophila melanogaster*. This cascade occurs in the microvilli of the rhabdomere, a structure within the cell body of the photoreceptor cell. The opsin interacts with a G alpha q that activates phospholipase C (PLC) initiating a phosphoinositide cascade that culminates in depolarisation of the photoreceptor cell. (**B**) Ciliary phototransduction in *Homo sapiens*. This cascade occurs in a specialised cilium of the photoreceptor cell. The opsin activates the G alpha of transducin that in turn activates phosphodiesterase 6 with consequent cascade that causes the hyperpolarization of the photoreceptor cell. In rod photoreceptors, the opsin and the other membrane proteins, with the exception of the ion channels, are in the membrane of the disk as depicted here. In cone photoreceptors, whilst the components and the cascade are the same, all membrane components are in the cell membrane (not depicted here). The pathways are based primarily on the Kegg maps ko04745 (rhabdomeric) and ko04744 (ciliary). Additional references were (Hardie and Juusola 2015) for *D. melanogaster* phototransduction and (Lamb 2020) for *H. sapiens*.

Protein components are coloured in red (rhabdomeric pathway) or green (ciliary pathway), ions and other non-protein molecules are represented by small grey circles. Lines between components indicate physical interaction, normal arrows between components indicate activation, normal arrows through channels indicate passage of ions, inhibitory arrows indicate inactivation, dotted arrows indicate movement/ transition towards, +p indicates phosphorylation, -p indicates de-phosphorylation, ? indicates unclear mechanism.

The two phototransduction pathways occur in different subtypes of photoreceptor cells (PRCs). Rhabdomeric PRCs utilise the phosphoinositide pathway while the ciliary PRCs use the phosphodiesterase 6 pathway (Arendt 2003). Both cell types occur throughout Metazoa (Horridge 1964; Hattar et al. 2002; Arendt 2003; Nordström et al. 2003; Arendt et al. 2004; Kozmik et al. 2008; Passamaneck et al. 2011; Ullrich-Lüter et al. 2011; Jékely et al. 2015; Tamm 2016; von Döhren and Bartolomaeus 2018; Picciani et al. 2018; Valencia et al. 2021). The homology of the two photoreceptor cell types is still under debate, as is the question of the ancestral state in the ancestor to all animals (Arendt 2008; Arendt et al. 2016). Traditionally, the identification of PRCs in animals was based on morphological studies. However, there is now an increasing understanding that PRCs are characterised by a distinct molecular profile, including both the molecular machinery required for phototransduction and a core set of regulatory genes crucial for their cell-type identity (Arendt 2003; Arendt 2008; Arendt et al. 2016; Valencia et al. 2021). The availability of single-cell RNA sequencing technologies have resulted in a growing number of datasets that allows us to explore the presence of these cell-type molecular profiles in diverse organisms.

In this chapter I explored the evolutionary history of the molecular components essential for vision by first reconstructing the evolution of the genes involved in the two major phototransduction pathways, and then identifying PRC-like cell-types throughout animals and comparing their genetic profiles.

Table 3.1. All phototransduction components with respective gene and protein names. Common components are listed for both *Drosophila melanogaster* and *Homo sapiens*. Rhabdomeric components are listed for *D. melanogaster* and ciliary components are listed for *Homo sapiens*. The gene and protein names are based on FlyBase, GeneCards and UniProt.

	Component	*Drosophila melanogaster*		*Homo sapiens*	
	(Gene Family)	**Gene Name(s)**	**Protein Name(s)**	**Gene Name(s)**	**Protein Name(s)**
Common	Opsin	ninaE Rh 2 to 7	Opsin Rh1 Opsins Rh2 to 7	OPN1LW OPN1SW OPN1MW RHO	Long-wave-sensitive opsin 1 Short-wave-sensitive opsin 1 Medium-wave-sensitive opsin 1 Rhodopsin
	G beta	Gbeta76C	Gbeta76C (or Gbe)	GNB 1 to 4	G protein subunit beta 1 to 4
	G gamma	Ggamma30A	Ggamma(e)	GNGT1 GNGT2	G protein G(T) subunit gamma-T1 G protein G(T) subunit gamma-T2
	Calmodulin	Cam	Calmodulin (or CaM)	CALM 1 to 3	Calmodulin 1 to 3
	Arrestin	Arr1 Arr2	Phosrestin-2 (or Arrestin-1) Phosrestin-1 (or Arrestin-2)	ARR3 SAG	Arrestin-C S-arrestin
	GRK	Gprk1	GPCR kinase 1	GRK1 GRK7	Rhodopsin kinase GRK1 Rhodopsin kinase GRK7

	Component	*Drosophila melanogaster*	
	(Gene Family)	**Gene Name(s)**	**Protein Name(s)**
Rhabdomeric	G alpha q	Galphaq	G protein alpha q subunit
	PLC	norpA	Phosphoinositide phospholipase C-beta
	PKC	inaC	Protein kinase C, eye isozyme (or Eye-PKC)
	INAD	inaD	Inactivation-no-after-potential D protein
	MYO3	ninaC	Neither inactivation nor afterpotential protein C
	Actin	Act5C	Actin-5C
	TRP C	trp trpl	Transient receptor potential protein Transient-receptor-potential-like protein
	IP3R-A	Itpr	Inositol 1,4,5-trisphosphate receptor (or IP3R)
	CamKII	CaMKII	Calcium/calmodulin-dependent protein kinase
	DAGL	inaE	Inactivation no afterpotential E
	PPEF	rdgC	Serine/threonine-protein phosphatase rdgC (or Retinal degeneration C protein)

	Component	*Homo sapiens*	
	(Gene Family)	**Gene Name(s)**	**Protein Name(s)**
Ciliary	G alpha i	GNAT1 GNAT2	G protein G(t) subunit alpha 1 G protein G(t) subunit alpha 2
	PDE6 A/B/C	PDE6A PDE6B PDE6C	Rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha Rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta Cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'
	PDE6 G/H	PDE6G PDE6H	Retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit gamma Retinal cone rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit gamma
	GC2	GUCY2D GUCY2F	Retinal guanylyl cyclase 1 (or GC2D) Retinal guanylyl cyclase 2 (or GC2F)
	GCAP	GUCA1A GUCA1B GUCA1C	Guanylyl cyclase-activating protein 1 (GCAP1) Guanylyl cyclase-activating protein 2 (GCAP2) Guanylyl cyclase-activating protein 3 (GCAP3)
	CNG	CNGA 1 to 4 CNGB 1 and 3	cGMP-gated cation channel alpha 1 to 4 Cyclic nucleotide-gated cation channel beta 1 and 3
	NCKX	SLC24A 1, 2 and 4	Sodium/potassium/calcium exchanger 1, 2 and 4 (or NCKX1,2 and4)
	Recoverin	RCVRN	Recoverin
	RGS9	RGS9	Regulator of G-protein signaling 9
	RGS9BP	RGS9BP	Regulator of G-protein signaling 9-binding protein (RGS9BP or RGBP) or RGS9-anchoring protein (R9AP)
	GNB5	GNB5	G protein subunit beta 5

Results and Discussion

Extended gene families of phototransduction components are generally broadly distributed throughout eukaryotes.

To study the evolution of the phototransduction cascades, I first investigated the distribution of each phototransduction component gene family in 86 eukaryotic species (Table 3.2). I focused on non-bilaterian animals (25 species) and closest relatives of animals (8 choanoflagellates and 5 other holozoans), but also included a balanced sampling of all major eukaryotic groups to account for the potential ancient origin of some of the gene families (see Supplementary Table S3.1 with source information). The phototransduction components examined were based primarily on the Kegg maps ko04745 (*D. melanogaster* rhabdomeric cascade) and ko04744 (*H. sapiens* ciliary cascade). The data mining was carried out with a combination of sequence similarity and protein motif analyses (details are described in the Methods section). I then constructed maximum likelihood phylogenetic trees followed by gene tree to species tree reconciliations for each gene family (see Methods for details). Gene tree to species tree reconciliations were performed both with ctenophore-first and sponge-first scenarios and comparison of total number of events (duplications and losses) revealed that, overall, there were no major differences between the two scenarios (Supplementary Table S3.2).

Most gene families examined were broad, therefore, within each gene family I focused on identifying the sub-group containing *D. melanogaster* and/or *H. sapiens* genes known to function in the phototransduction cascades. The objective was understanding whether early-branching animals and non-animal species might possess genes that could perform in phototransduction, and in many cases this might include non-orthologous but related genes, therefore, I expanded the definition of the group of interest to include a broader set of genes within an orthogroup of interest. While the specific orthogroup of interest is often present only within animals or in sister-groups to animals (Figure 3.2), closely related sub-groups were present in the next related species, and when considering the whole extended gene family, the distribution would often span Eukarya. This adds an

extra layer of detail to our knowledge of when exactly the functional phototransduction pathways might have originated.

Clades							Species	% Complete BUSCOs (tot) (eukaryota_odb10)
Amorphea	Obazoa	Opisthokonta	Holozoa	Metazoa	Bilateria	Arthropoda	*Drosophila melanogaster*	100.00%
							Calanus glacialis	96.80%
							Daphnia pulex	97.70%
							Strigamia maritima	92.60%
						Nematoda	*Caenorhabditis elegans*	97.70%
							Pristionchus pacificus	80.40%
							Loa loa	96.10%
						Tardigrada	*Ramazzotius varieornatus*	92.50%
						Mollusca	*Octopus bimaculoides*	92.90%
							Lottia gigantea	96.50%
						Brachiopoda	*Lingula unguis*	98.40%
						Annelida	*Capitella teleta*	94.20%
							Helobdella robusta	97.30%
						Bryozoa	*Bugula neritina*	54.50%
						Vertebrata	*Homo sapiens*	100.00%
							Mus musculus	100.00%
							Danio rerio	99.60%
							Eptatretus burgeri	85.90%
						Urochordata	*Ciona intestinalis*	97.30%
						Cephalochordata	*Branchiostoma belcheri*	96.90%
						Echinodermata	*Acanthaster planci*	91.40%
							Strongylocentrotus purpuratus	96.00%
						Hemichordata	*Saccoglossus kowalevskii*	94.10%
					Cnidaria		*Acropora digitifera*	42.80%
							Acropora tenuis	27.80%
							Astreopora sp	69.80%
							Porites australiensis	81.60%
							Fungia scutaria	84.40%
							Montastraea cavernosa	80.40%
							Madracis auretenra	68.30%
							Stylophora pistillata	85.10%
							Anthopleura elegantissima	62.00%
							Nematostella vectensis	93.40%
							Gorgonia ventalina	56.50%
							Clytia hemisphaerica	84.70%
							Hydra magnapapillata	70.90%
							Aurelia sp	67.40%
					Placozoa		*Trichoplax adhaerens*	96.10%
							Hoilungia hongkongensis	96.80%
					Porifera		*Amphimedon queenslandica*	92.50%
							Haliclona tubifera	75.30%
							Ephydatia muelleri	93.40%
							Stylissa carteri	43.60%
							Leucosolenia complicata	94.90%
							Sycon ciliatum	97.30%
							Oscarella pearsei	94.90%
					Ctenophora		*Mnemiopsis leidyi*	83.60%
							Pleurobrachia bachei	47.50%
				Choanoflagellata			*Acanthoeca spectabilis*	93.30%
							Helgoeca nana	94.10%
							Diaphanoeca grandis	92.20%
							Didymoeca costata	94.90%
							Choanoeca perplexa	92.10%
							Monosiga brevicollis	78.80%
							Mylnosiga fluctuans	93.30%
							Salpingoeca kvevrii	94.90%
				Ichthyosporea			*Amoebidium parasiticum*	89.90%
							Sphaeroforma arctica	63.20%
							Sphaerothecum destruens	68.60%
				Filastrea			*Capsaspora owczarzaki*	93.70%
				Pluriformea			*Corallochytrium limacisporum*	90.20%
			Nucletmycea (Holomycota)	Fungi			*Fusarium oxysporum*	97.30%
							Saccharomyces cerevisiae	93.30%
							Ustilago maydis	98.80%
							Mortierella elongata	98.40%
							Rhizopus microsporus	97.30%
							Spizellomyces punctatus	94.90%
				Rotosphaerida			*Parvularia atlantis*	75.70%
							Fonticula alba	71.00%
		Apusomonadida					*Thecamonas trahens*	81.60%
		Breviata					*Pygsuia biforma*	62.40%
	Amoebozoa						*Dictyostelium discoideum*	94.10%
							Vermamoeba vermiformis	89.00%
							Cunea sp	76.50%
	CRuMs	Collodictyonidae					*Diphylleia rotans*	90.60%
		Rigifilida					*Rigifila ramosa*	86.30%
Diaphoretickes	Archaeplastida			Rhodelphis			*Rhodelphis marinus*	89.80%
				Rhodophyta			*Galdieria sulphuraria*	77.30%
				Chloroplastida			*Arabidopsis thaliana*	99.60%
	Cryptista			Cryptophyceae			*Guillardia theta*	79.60%
	Haptista			Haptophyta			*Phaeocystis globosa*	75.30%
	Sar			Rhizaria			*Plasmodiophora brassicae*	89.00%
				Stramenopiles			*Phytophthora infestans*	90.20%
				Alveolata			*Colponemidia sp*	93.30%
	Excavata			Euglenozoa			*Euglena longa*	83.50%
				Heterolobosea			*Pharyngomonas kirbyi*	85.10%

>90%
>80%
>70%
>60%
>50%
<50%

Table 3.2. List of eukaryotic species used for the phylogenetic analysis of phototransduction gene families. The respective percentages of total complete BUSCO genes are indicated. BUSCO was conducted using species proteomes versus the eukaryota_odb10 database.

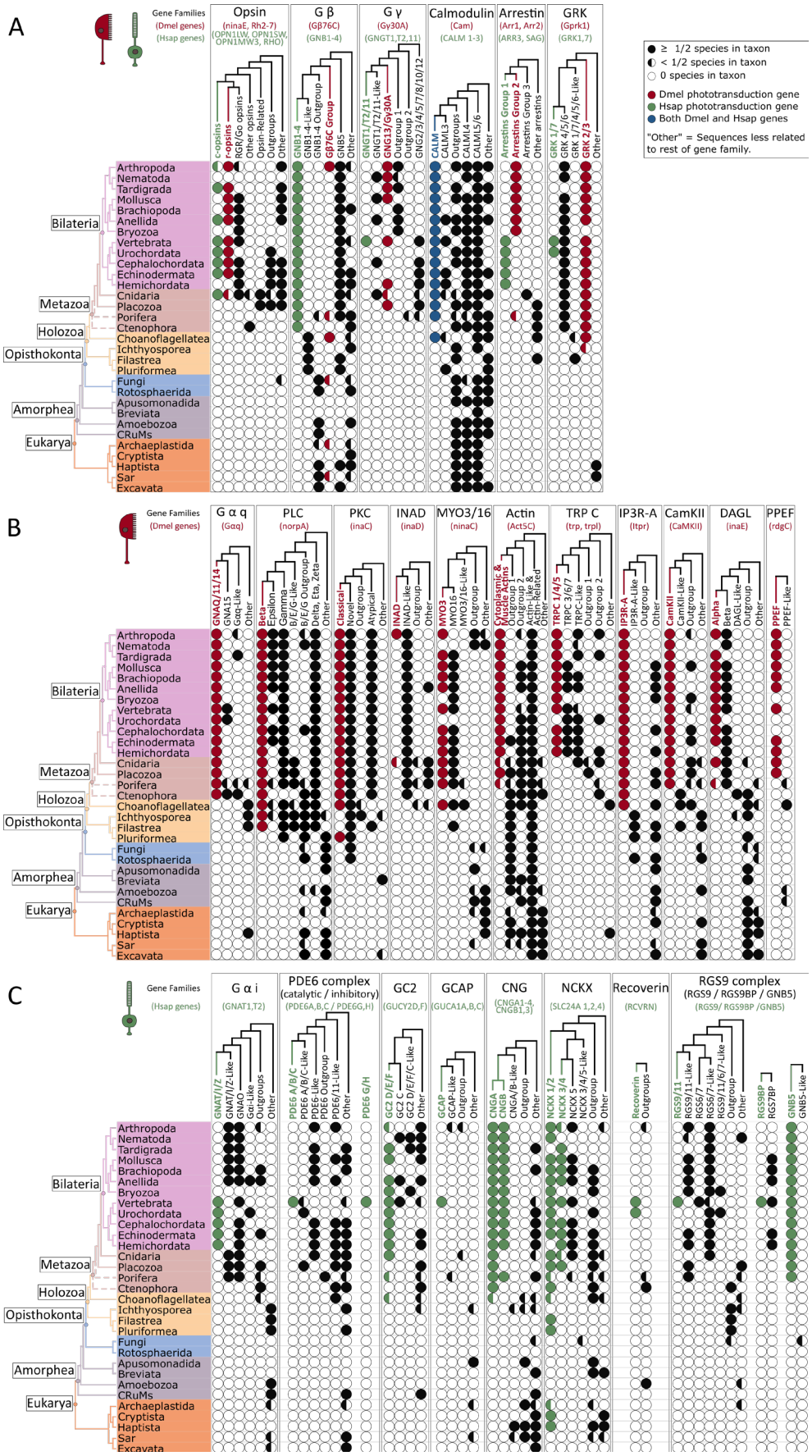


Figure 3.2. Evolutionary history of phototransduction components gene families and distribution across Eukarya. We reconstructed the evolution of each gene family for all common (**A**), rhabdomeric-specific (**B**) and ciliary-specific (**C**) components and we mapped their distribution across all major groups of Eukarya. For each gene family, we obtained a gene tree based on maximum likelihood phylogenetic

trees and gene tree to species tree reconciliations. Most gene families examined were broad, therefore, within each gene family tree we highlight the branch containing the *D. melanogaster* and/or *H. sapiens* gene that is known to function in the phototransduction pathway. When mapping the presence/absence of the phototransduction components throughout the tree of eukaryotes, we distinguish for each gene family whether the presence refers to the specific orthogroup of interest or to any of the other related sub-lineages within the broad gene family. While the specific-orthogroup of interest is often present only within animals or in sister-groups to animals, we detected numerous cases in which organisms more distantly related to animals possessed related genes within the broad gene family.

Common phototransduction components

There are six gene families that are common amongst the two phototransduction pathways (Table 3.1). Within these families, the orthogroups of interest are the ones that contain either the human or the fly gene (or both). These orthogroups of interest are present either strictly in animals, like in the case of opsins, G gamma and arrestin, or within Holozoa (G beta, calmodulin, GRK) (Figure 3.2A). However, if we consider the extended gene family, then G beta, calmodulin and GRK span all Eukarya, arrestin is present up to Holozoa, and only opsin and G gamma remain animal-specific. While for opsin this reflects the expected scenario (Fleming et al. 2020), for G gamma, an animal-oriented definition of the gene family (e.g. during protein motif filtering, see Methods) may have resulted in the exclusion of non-animal G gamma-types. Indeed, G gamma is the least studied subunit of the G protein and its subtypes outside of animals are not well characterised (Krishnan et al. 2015). As the other two subunits of the G protein (G beta and G alpha) are present outside of animals, it is possible that a non-animal G gamma-type exists but was not detected here. Of note, in Figure 3.2, the G alpha family is divided into two subfamilies corresponding to those used in fly rhabdomeric (Gq family) (Hardie and Juusola 2015) and vertebrate ciliary (Gi/o family) (Lagman et al. 2012) phototransduction, as these are the two pathways that were used here as reference. However, I also conducted a comprehensive analysis of the full G alpha family that includes other important subfamilies such as Gs (see supplementary reconciliation files on GitHub).

Rhabdomeric-specific phototransduction components

Within the gene families of the rhabdomeric-specific components (Figure 3.2B), the orthogroup of interest is animal-specific for seven out of eleven gene families, with the remaining four families being holozoan-specific. It is therefore of striking contrast that

the extended gene families are all present throughout Eukarya, except for INAD (inactivation no afterpotential D) that appears to be restricted to animals and choanoflagellates. The presence outside of Holozoa is perhaps questionable also for G alpha q and TRP C, however, overall rhabdomeric extended gene families appear to be ancient.

Ciliary-specific phototransduction components

The majority of the ciliary-specific orthogroups of interest (Figure 3.2C), are also animal-specific (eight out of eleven). Two components are present also in Holozoa, while only NCKX, a sodium-calcium-potassium exchanger involved in numerous other pathways (Altimimi and Schnetkamp 2007), is present throughout Eukarya. The situation is dramatically different if you consider the extended gene families, as in this case nine families are present in Eukarya and only two remain animal-specific. Of further note for the ciliary components, in contrast to common and rhabdomeric components that within animals are distributed in all or most phyla, the ciliary components often have a patchy presence also within animals, with vertebrates being the only group that contains all the gene families. This indicates that some of the components of the ciliary pathway used as a reference are likely vertebrate innovations, while other components are more ancient and represent the core part of the cascade. An example of this can be seen for the PDE6 complex. The alpha/beta subunits belong to the same protein family and constitute the essential catalytic subunits of the complex, while the gamma subunits are accessory inhibitory subunits that have been described only in vertebrate PDE6 (Lagman et al. 2016; Lamb 2020). Here, my result confirms the notion that PDE6 gamma subunits are a vertebrate novelty.

Patterns of major duplication, speciation and loss events clarify gene family expansions.

The approach of reconciling the gene trees to the species tree not only allowed to define the orthogroups of interest, but also revealed the specific patterns of duplication, speciation and loss events that characterise the lineage of the orthogroup of interest and

all other lineages in the gene family. Here, I discuss the key findings for a few gene families of particular interest.

GPCR Kinases: an ancient family that expands in Metazoa

An interesting case amongst the common phototransduction components is that of the G-protein-coupled receptor kinases (GRK) (Figure 3.3A). This family has an ancient origin with presence in some distantly related eukaryotes, however, it is characterised by a series of key duplications just prior to and at the base of animals that gave rise to the various sub lineages of interest for either rhabdomeric or ciliary phototransduction (Figure 3.3A).

In photoreceptor cells, the GRKs are essential for the inactivation phase of phototransduction. The light-activated visual pigment is capable of activating hundreds of G proteins (Shichida and Matsuyama 2009). To avoid the signal to continue long after the original light stimulus occurred, the visual pigment must be shut-off (Wang and Montell 2007; Lamb et al. 2018). After shut-off, photoreceptors have to recover their pre-illumination state and the quicker this occurs, the more they can adjust to rapidly changing lighting conditions (Orban and Palczewski 2016). GRKs, protein kinases of the serine/threonine protein kinases superfamily, phosphorylate target GPCRs facilitating the binding of arrestin to the GPCR (Mushegian et al. 2012; Orban and Palczewski 2016). The arrestin-capped GPCR is blocked from interacting with its G-protein. Therefore, GRKs initiate the desensitisation of GPCRs and deactivation of GPCR signalling (Gurevich and Gurevich 2016; Orban and Palczewski 2016).

In vertebrates there are seven GPCR kinases (GRK 1-7), and the ones involved in phototransduction shut-off are GRK1 (rods) and GRK7 (cones) (Lamb et al. 2018; Lamb 2020). The fruit fly *Drosophila melanogaster* possesses two GRK genes, Gprk1 and Gprk2. The GRK involved in phototransduction shut-off is Gprk1, which is more closely related to GRK 2/3 (Lee et al. 2004; Wang and Montell 2007). Overall within Metazoa, the GRK family is split into two major clades: one clade includes GRK 2 and 3; while the other contains all other GRKs and in turn is composed of two subgroups, one with GRK 1 and 7 and the other with GRK 4, 5, and 6 (Mushegian et al. 2012). An extensive phylogenetic analysis of GRKs (Mushegian et al. 2012) previously found that GRKs are an ancient family that arose well before Metazoa. In that study, the authors concluded that the GRK family underwent a first split into GRK 2/3 type and GRK 1/7+4/5/6 type at some point before the advent of animals within the history of opisthokonts. Further expansions occurred later within animals, likely to reflect the greater need for rapid signalling to adapt to the surrounding environment (Mushegian et al. 2012).

With my much broader set of eukaryotic lineages examined, the focus on early-branching animals and sister groups of animals, and the gene tree to species tree reconciliation, I was able to expand our knowledge of the evolution of the GRK family adding further details compared to the work of Mushegian et al. (Mushegian et al. 2012). In accordance with previous results, the duplication that gives rise to the GRK 1 and 7 sub-groups is at the split between urochordates and vertebrates (see supplementary files with the full reconciliation for GRK on GitHub). The GRK 4/5/6 sub-groups all derive from two subsequent duplications at the base of jawed vertebrates. Interestingly, the split between GRK 1/7 and GRK 4/5/6 appears to be much more ancient than expected as it derives from a gene duplication at the base of Metazoa. This holds true in both ctenophore-first and sponge-first scenarios (Figure 3.3A). Whilst the GRK 4/5/6 lineage is widespread throughout animals, the GRK 1/7 lineage seems to have been lost in all animal groups except in Olfactores (urochordates and vertebrates) and potentially in ctenophores, according to the sponge-first scenario only (Figure 3.3A).

The duplication that gave rise to the split between GRK 2/3 and GRK 1/7+4/5/6 occurred at the base of Holozoa (Figure 3.3A). Therefore, the closest relatives to Metazoa inherited both lineages, as previously proposed (Mushegian et al. 2012). Although my results too show that several holozoans lost either one or the other lineage as described previously, my larger taxonomic sampling allowed me to clarify that at least within choanoflagellates, both lineages were originally present, contrary to what previously thought (Mushegian et al. 2012).

Finally, outside of Holozoa, GRKs are not present in other opisthokonts (e.g., fungi) nor in any other Amorphea group (e.g., Amoebozoa). An orthologous lineage to the GRKs 1-7 is instead present in the other major eukaryotic branch, the Diaphoretickes (Figure 3.3A). However, the presence is limited to a small subset of groups, namely the SAR and Haptophyta (see supplementary reconciliation files).

Gene tree to species tree reconciliations under either ctenophore-first or sponge-first scenarios provided the same overall results, with one minor exception: in ctenophore-first scenario, the GRK 1/7 lineage is present only in Olfactores and the ctenophore branch includes only GRK 2/3 and GRK 4/5/6; instead in the sponge-first scenario, the GRK 1/7 lineage is present also in the ctenophore branch, that has lost GRK 4/5/6 (but retained GRK 2/3) (Figure 3.3A).

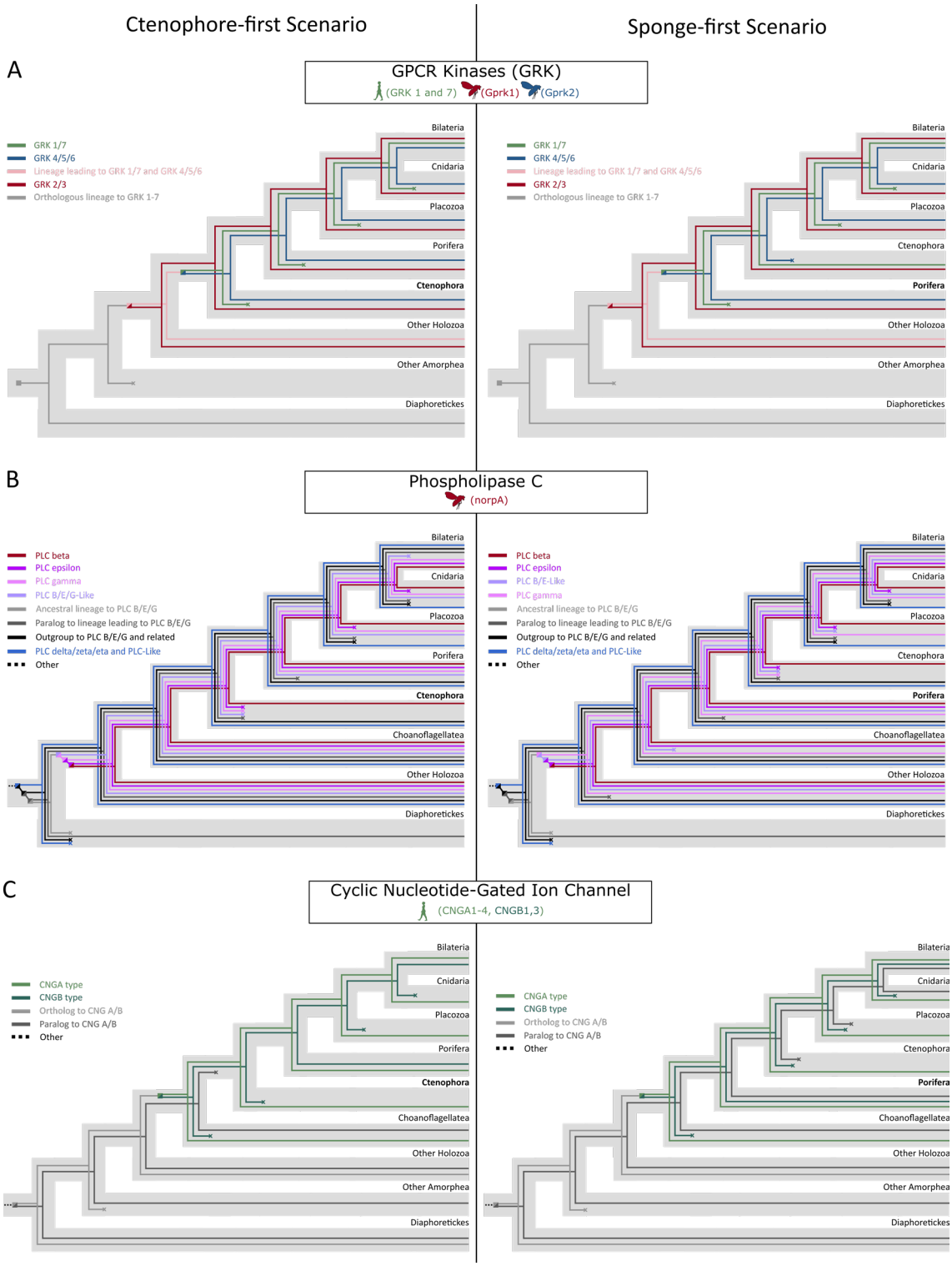


Figure 3.3. Major events of duplication, speciation, and losses for three phototransduction gene families of interest. Reconciliations were constructed under both ctenophore-first and sponge-first scenarios and no major differences were found. (**A**) GPCR Kinases (GRK) are important for the shut-off of light response in both rhabdomeric and ciliary phototransduction. The gene family has an ancient eukaryotic origin, however the key duplication events that gave rise to the diversity of the family present in animals, occurred just prior to animals and at the base of animals. The lineage that gives rise to the *Drosophila melanogaster* gene Gprk1 that is used in rhabdomeric phototransduction derives from a duplication at the base of Holozoa. While a duplication at the base of animals gave rise to the lineage that includes the human GRK1 and GRK7 involved in ciliary phototransduction. (**B**) The phospholipase C **(**PLC) is important for the initial steps of the rhabdomeric phototransduction. It is a very broad family of enzymes that includes

many subgroups. The *Drosophila* gene NorpA involved in phototransduction is a PLC type beta. This lineage, like most others in the family, derived from a duplication at the base of Holozoa. (**C**) The cyclic nucleotide gated ion channels **(**CNG) are responsible for the hyperpolarization of vertebrate photoreceptor cells at the end of the signal cascade. It is again a very ancient family and the two subunits, alpha and beta, that compose vertebrate CNG channels originated from a duplication at the split between choanoflagellates and animals.

Phospholipase C: Holozoan origin of the beta subgroup from an ancient eukaryotic family

As mentioned, most rhabdomeric gene families have an ancient origin. An example of a gene family with an extensive repertoire of sub lineages deriving from very ancient gene duplications is the family of phospholipases of type C (PLC) (Figure 3.3B). PLCs are a broad family of enzymes that catalyse the hydrolysis of the phospholipid PIP2 into DAG and IP3, that both function as second messengers (Suh et al. 2008). In *Drosophila* a PLC of type beta is the one used in phototransduction. During the phototransduction cascade, IP3 interacts with its receptor (IP3R) on the endoplasmic reticulum, causing the release of calcium, and DAG goes on to activate the eye-specific protein kinase C (PKC) that is involved in the deactivation of the visual cascade (Wang and Montell 2007; Hardie and Juusola 2015).

Although PLCs have been described throughout Eukarya (Tsutsui et al. 1995; Koyanagi et al. 1998; Rebecchi and Pentyala 2000; Mikami 2014; Wang et al. 2020), few studies have investigated their evolution. In mammals there are 6 subgroups of PLCs: beta; gamma; delta; epsilon; zeta; and eta (Suh et al. 2008). Candidate beta-type and gamma-type PLCs have been cloned in the sponge *Ephydatia fluviatilis* and a delta-like in the cnidarian *Hydra magnipapillata* (Koyanagi et al. 1998). While the PLCs in fungi (and plants) have been described as similar to delta-type (Rebecchi and Pentyala 2000). A comprehensive phylogenetic analysis of the family is lacking.

The data mining recovered for Human, the 13 known PLCs belonging to the 6 subgroups plus two inactive PLC-Like sequences; and for *Drosophila*, the PLC beta used in phototransduction, encoded by NorpA, plus two other PLCs: PLC21C and small wing (sl). Gene tree to species tree reconciliation revealed that *Drosophila* NorpA arises from a duplication in the stem group of Cnidaria and Bilateria, and that from the same duplication arises Human PLCbeta4 (see supplementary files with the full reconciliation for PLC on GitHub). Instead, *Drosophila* PLC21C is more related to Human

PLCbeta1/2/3, and their lineage originates with a duplication at the base of Metazoa. A prior duplication at the same species node is the one that separates the PLC21C + PLCbeta1/2/3 on the one hand from the NorpA + PLCbeta4 on the other. These duplication patterns are consistent between ctenophore-first and sponge-first scenarios. Several additional duplications for the PLC beta lineage also occur at the base of Metazoa (in both ctenophore-first and sponge-first scenarios), indicating that PLC beta underwent a great expansion at the base of Metazoa (see supplementary reconciliation files). The origin of the PLC beta lineage is from a duplication at the base of Holozoa where its paralog lineage is the PLC epsilons (Figure 3.3B). At the same species node, a previous duplication gave rise to the PLC beta/epsilon lineage on the one hand and the PLC gamma on the other. The position of PLC epsilon as sister group to PLC beta is recovered with both ctenophore-first and sponge-first scenarios. This is a novel insight into the evolution of PLC subfamilies, as PLC beta has been considered to be related to gamma and delta (Rebecchi and Pentyala 2000), while here my results show that its closest relative seems to be epsilon. My results show that PLC beta/epsilon/gamma are more related to each other than to the other PLCs including PLC delta (Figure 3.3B). This clarification can be crucial, especially when trying to identify possible candidate genes involved in a putative rhabdomeric-like phototransduction pathway in non-model organisms such as non-bilateria. Tracing backwards the lineage of PLC beta/epsilon/gamma, uncovers that it originates from a duplication at the root of eukaryotes (Figure 3.3B). Here at this species node, there are multiple other duplications, including the one that gives rise to the lineage of all the other subgroups of PLCs (delta, zeta, eta) known in mammals. These major subgroup relationships remain consistent between ctenophore-first and sponge-first scenarios (Figure 3.3B).

Cyclic Nucleotide Gated Ion Channels: ancient origin of alpha and beta subtypes

Amongst the ciliary phototransduction components the cyclic nucleotide gated ion channels (CNGs) gene family is one of the ones with the broadest distribution across Eukarya (Figure 3.2C and Figure 3.3C). CNGs belong to the broad family of voltage-gated ion channels (Anderson and Grenberg 2001) and function in response to the binding of cyclic nucleotides. They are non-selective cation channels through which the passage of Ca2+ ions is of particular importance for the excitation of sensory cells (Kaupp and Seifert 2002).

During phototransduction the drop of cyclic GMP, caused by its hydrolysis by phosphodiesterase (PDE), induces the closure of CNG channels which in turn causes the hyperpolarization of the photoreceptor cell. Apart from this role in the activation of phototransduction, CNG channels are also involved in the Ca2+-feedback regulation of the cascade and thus in photoreceptor light adaptation (Kaupp and Seifert 2002).

The ion channel complex is composed of two groups of subunits, alpha and beta. Jawed vertebrates possess six genes encoding for CNG subunits: CNGA1-4 encode for four alpha subunits while CNGB1 and CNGB3 encode for beta subunits (Kaupp and Seifert 2002, Lamb 2020). The ion channel complex consists in the combination of four subunits around a pore. Native rod channels consist of three alpha1 (CNGA1) and one beta1 (CNGB1) subunits, while cone channels comprise two alpha3 (CNGA3) and two beta3 (CNGB3) subunits. Subunits alpha2 (CNGA2) and alpha4 (CNGA4) together with beta1 (CNGB1) are instead used in CNG channels of olfactory receptor neurons. Phylogenetic and gene synteny analyses led (Lamb 2020) to the reconstruction that the gene lineages of alpha and beta subunits derived from a duplication that occurred before the split of protostomes and deuterostomes (Lamb 2020). Likewise, CNGA4 split from the other branch of CNGA that later gave rise to CNGA1-3, prior to the protostome-deuterostome split. The authors speculate that the ancestral CNG channel was composed of two alpha and two beta subunits (Lamb 2020).

Outside of vertebrates, homologs to the CNG genes have been found in the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* and the horseshoe crab *Limulus polyphemus*, where likely they are involved in chemosensation (Kaupp and Seifert 2002). Amongst early branching animals, CNGs have been found in the cnidarian *Hydra magnipapillata* where it is implicated in phototransduction (Plachetzki et al. 2010). CNG channels are in fact not confined to animals as they are present also in plants (Saand et al 2015) and prokaryotes (Brams et al 2014, Napolitano et al 2021). However, while much attention has been given to the evolution of the CNG genes within and at the base of vertebrates, not much is known about the ancient evolutionary history of this gene family and the relationship between animal and non-animal CNG lineages.

The phylogenetic analysis and gene tree to species tree reconciliation of the CNG family revealed that the alpha and beta gene lineages derive from a gene duplication within the ancestor of choanoflagellates and animals (Figure 3.3C). This is independent of the

ctenophore-first or sponge-first scenario. Although it was already hypothesised that this gene duplication was ancient (Lamb 2020), it had not yet been clarified when it had occurred precisely.

According to my reconstructions, while the alpha lineage seems to be present in all major animal groups and choanoflagellates, the beta lineage seems to be present only in Bilateria and sponges (Figure 3.3C). The orthologous lineage to the CNG alpha/beta lineage is present in other holozoan species and in some Diaphoretickes. More distantly related CNG genes are present throughout Eukarya, but not in animals, according to the ctenophore-first reconciliation. While in the sponge-first reconciliation, this group of less related CNGs appears to be present also in Porifera, Cnidaria and Bilateria.

Identification of putative photoreceptor cells throughout animals.

To understand the origin and early evolution of vision, we must understand not only when all the individual phototransduction genes evolved, but also when they became co-expressed within the same cell-type. My detailed analysis of phototransduction gene family evolution with clarifications of the relationships amongst sub lineages, allowed me to compile a list of best candidate phototransduction genes for every species. These were used as markers to identify candidate photoreceptor cells (PRCs) from the available single-cell RNA sequencing data of a variety of animal species. I focused the investigation on twelve species that spanned Metazoa with particular emphasis on early-branching animals. *Drosophila melanogaster* was used as representative of rhabdomeric PRCs; *Homo sapiens* and *Mus musculus* as representatives of ciliary PRCs. The urochordate *Ciona intestinalis* and the sea urchin *Strongylocentrotus purpuratus* served as invertebrate representatives of the deuterostome clade, useful to "bridge" the gap between the two primary model organisms—fly and human. Finally, amongst non-bilaterians I investigated the cnidarians *Hydra vulgaris*, *Clytia hemisphaerica*, *Stylophora pistillata* and *Nematostella vectensis*; the placozoan *Trichoplax adhaerens*, the sponge *Amphimedon queenslandica* and the ctenophore *Mnemiopsis leidyi*. A comprehensive list of scRNAseq data sources and sample details for each species are in Table 3.3.

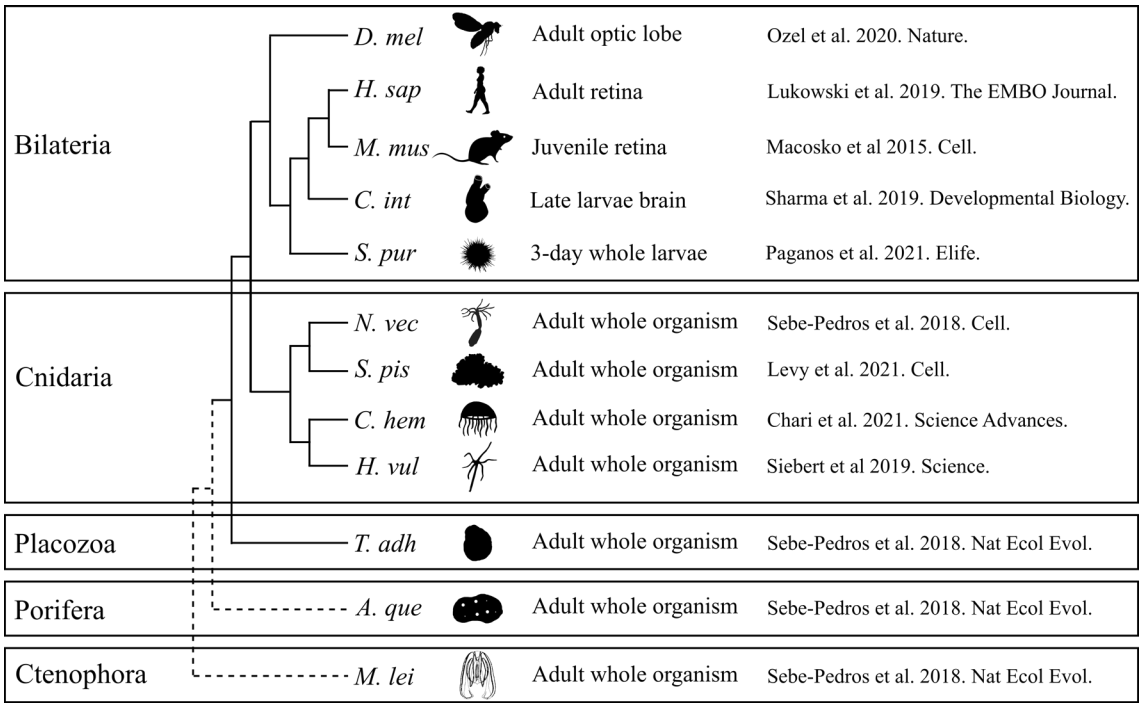


Table 3.3. Datasets used for the single cell analyses. Single cell RNA sequencing data from 12 species were used for the single cell analyses. Phylogenetic relationships, as well as developmental stage, tissue type and source are indicated for each species. Focus was given on non-bilaterian species. Species silhouettes are images with CC0 1.0 Universal Public Domain Dedication licences obtained from https://www.phylopic.org/. Abbreviations: *D. mel*: *Drosophila melanogaster*; *H. sap*: *Homo sapiens*; *M. mus*: *Mus musculus*; *C. int*: *Ciona intestinalis*; *S. pur*: *Strongylocentrotus purpuratus*; *N. vec*: *Nematostella vectensis*; *S. pis*: *Stylophora pistillata*; *C. hem*: *Clytia hemisphaerica*; *H. vul*: *Hydra vulgaris*; *T. adh*: *Trichoplax adhaerens*; *A. que*: *Amphimedon queenslandica*; *M. lei*: *Mnemiopsis leidyi*.

In *D. melanogaster*, *H. sapiens* and *M. musculus* photoreceptor cells are well characterised. For *C. intestinalis*, *S. purpuratus* and some species of cnidaria, photoreceptors have been reported, while for most non-bilaterians evidence for photoreceptors is scant, like for ctenophores (Horridge 1964; Tamm 2016), or entirely unknown. Moreover, when searching for putative homologous cell-types to the PRCs in these species, it is uncertain whether they might possess a more rhabdomeric-like or ciliary-like profile. Therefore, I developed a pipeline (described in detail in the Methods section of this Chapter) to identify PRC-like "metacells" or cell states based on phototransduction gene expression.

The presence/absence of phototransduction genes, whether belonging to the best orthogroup or to another related lineage, provides some form of evidence to understand the diversity of PRC-like profiles amongst animals (Figure 3.4).

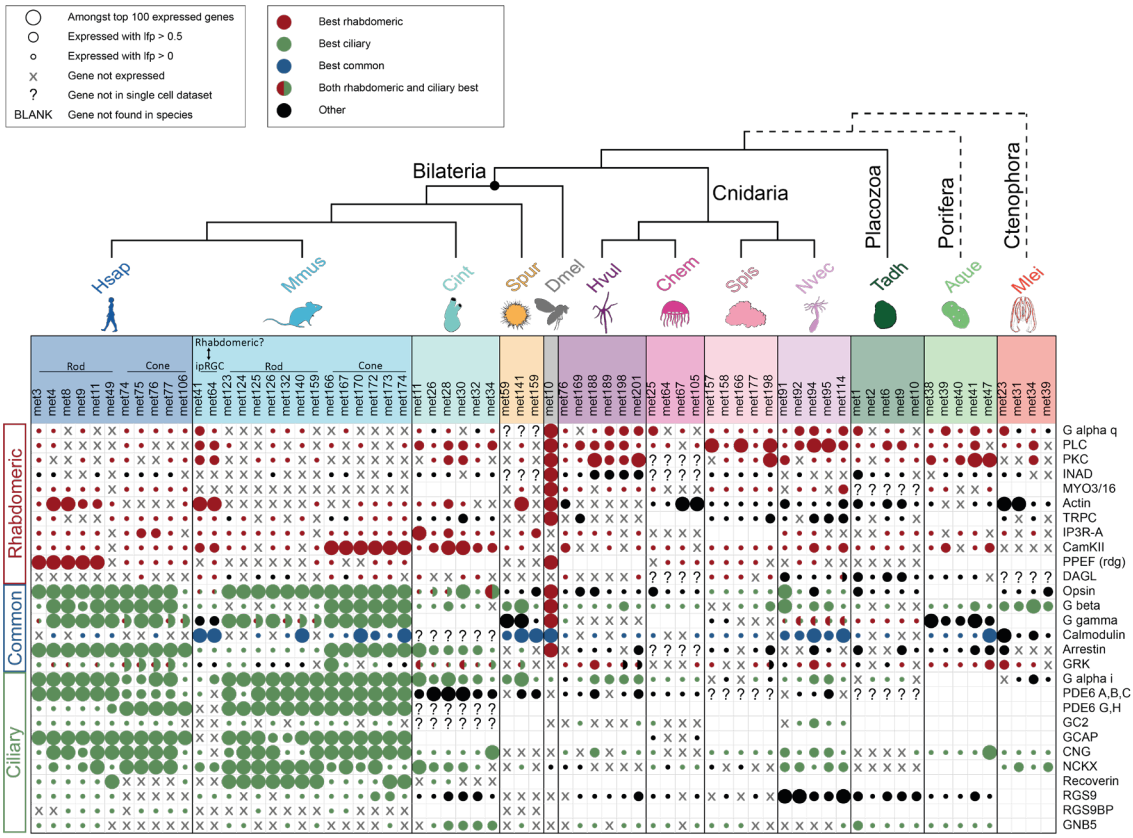


Figure 3.4. Expression of phototransduction genes in photoreceptor-like cells across animals. The single-cell RNA sequencing analysis identified putative PRC-like metacells across all the species examined, including all non-bilaterian phyla. Human and mouse ciliary PRCs express mainly ciliary type genes, but also some rhabdomeric type ones. Instead, *Drosophila* PRC expresses almost exclusively rhabdomeric type genes. Two candidate ipRGC (intrinsically photosensitive retinal ganglion cells) in mouse display a rhabdomeric-type profile. *Ciona intestinalis* metacells appear to have ciliary-like profiles. Outside of chordates, a large amount of phototransduction genes is either not present in the genome or not detected in the scRNAseq data, and, overall, most species have a mixture of rhabdomeric and ciliary genes expressed. Species silhouettes were modified from images with CC0 1.0 Universal Public Domain Dedication licences obtained from https://www.phylopic.org/. Abbreviations: *D. mel*: *Drosophila melanogaster*; *H. sap*: *Homo sapiens*; *M. mus*: *Mus musculus*; *C. int*: *Ciona intestinalis*; *S. pur*: *Strongylocentrotus purpuratus*; *N. vec*: *Nematostella vectensis*; *S. pis*: *Stylophora pistillata*; *C. hem*: *Clytia hemisphaerica*; *H. vul*: *Hydra vulgaris*; *T. adh*: *Trichoplax adhaerens*; *A. que*: *Amphimedon queenslandica*; *M. lei*: *Mnemiopsis leidyi*.

Ciliary and Rhabdomeric PRCs in model organisms

As expected, *D. melanogaster* PRC type expresses rhabdomeric phototransduction genes while *H. sapiens* and *M. musculus* express ciliary genes (Figure 3.4). The *Drosophil*a PRC metacell possesses a rhabdomeric-exclusive profile, with the only ciliary gene expressed being the G alpha of type i/o. All other ciliary genes were either not detected in the *Drosophila* genome or are not expressed in its PRC metacell. In contrast, *H. sapiens* and *M. musculus* metacells, while having a clear ciliary-oriented profile, still have a significant amount of rhabdomeric genes expressed, albeit at a lower level compared to ciliary genes. This could reflect the possibility that ciliary photoreceptor cells, whilst

using the ciliary pathway for phototransduction, may contemporarily employ rhabdomeric-like signalling either to modulate the phototransduction or, alternatively, to perform unrelated tasks, as previously proposed (Yau and Hardie 2009).

Furthermore, it has been proposed that melanopsin (OPN4) expressing cells in the vertebrate retina are the homologous cell type to rhabdomeric photoreceptor cells (Provencio et al. 2000; Hattar et al. 2002; Arendt 2003; Rollag et al. 2003; Fu et al. 2005). Specifically, a subclass of retinal ganglion cells, called intrinsically photosensitive retinal ganglion cells (ipRGC), are known to use OPN4 to mediate non-visual forming light sensing functions, e.g., circadian entrainment and pupillary light reflex (Hahn et al. 2023). The human OPN4 was not detected in the human retina single cell dataset used in this study, so no candidate rhabdomeric profile could be identified. Instead in the mouse dataset, two metacells were found to express the mouse OPN4 (metacells 41 and 64) (Figure 3.4). Both also expressed another ipRGC marker, the transcription factor EOMES (Hahn et al. 2023) (the list of all genes with their respective lfp values is available on GitHub). Interestingly, these two mouse metacells express a lot less ciliary genes compared to other mouse metacells. Metacell 64 in particular is missing some of the key genes involved in the ciliary pathway, including all the PDE6 subunits and the CNG channel. Additionally, metacells 41 and 64 are the mouse metacells that express the highest number of rhabdomeric genes, with metacell 64 expressing all except two rhabdomeric genes. These results suggest that based on phototransduction genes these two ipRGC metacells have a rhabdomeric profile.

C. intestinalis and S. purpuratus PRC metacells

The two deuterostome invertebrates examined here have both been reported to possess photoreceptor cells. The sea squirt *Ciona intestinalis* is known to possess a ciliary-type PRC (Eakin and Kuda 1970; Ryan et al. 2016). The sea urchin *Strongylocentrotus purpuratus* has been reported to have both rhabdomeric-type (Ullrich-Lüter et al. 2011) and ciliary-type (Valencia et al. 2021) PRCs. For both species the expression of phototransduction genes provided somewhat mixed results (Figure 3.4).

C. intestinalis metacells express both some rhabdomeric and some ciliary genes, with the common components being predominantly of ciliary type. However, many genes were either not found in the genome or not detected in the single cell data, so cannot be

assessed. If focusing on the opsins, then the majority of the metacells express only c-opsins, while some express contemporarily c-opsins and r-opsins. In this sense these results are consistent with the literature that has described a ciliary type PRC based on morphology (Eakin and Kuda 1970; Ryan et al. 2016). Whereas it was not possible to exclude or to suggest the possibility of the presence of a rhabdomeric-type PRC profile.

Similarly, in *S. purpuratus* several genes are missing either from the genome or from the single cell data (Figure 3.4). However, compared to *Ciona*, in the sea urchin there are also many genes that are present in the genome and the single cell data but that are not expressed in the PRC-like metacells. Of note I was only able to identify 3 PRC-like metacells in the sea urchin, likely because from all the opsins expressed in the genome, only two opsins were detected in the single cell data. For example, neither Sp-Opsin-4, an r-opsin described to be expressed in candidate rhabdomeric cells (Ullrich-Lüter et al. 2011), nor Sp-Opsin-3.2, a Go-opsin expressed in candidate ciliary cells (Valencia et al. 2021), were detected in the single cell dataset. The opsins that are in the single cell dataset (Sp-Opsin2 and Sp-Opn5L) are echinoderm-specific echinopsins (D'Aniello et al. 2015) that according to my phylogenetic analysis fall in the broad lineage of RGR/Go opsins (see supplementary files with the full reconciliation for opsins on GitHub). While they likely initiate a functioning phototransduction cascade, it is not certain whether it could be a rhabdomeric or ciliary pathway.

Photoreceptor-like metacells in non-bilateria

PRC-like in Cnidaria

Amongst all the non-bilaterian phyla, the Cnidaria are the only group in which there is clear evidence of the presence of photoreceptor cells (Piatigorsky and Kozmik 2004; Kozmik et al. 2008; Vöcking et al. 2022) and of which some components of the phototransduction cascade have been described (Plachetzki et al. 2010; Gornik et al. 2021). The results from my analysis revealed that although several phototransduction genes were missing in the genomes/transcriptomes and/or in the single cell data of cnidarian species, overall, this phylum seems to have the most complete repertoire of phototransduction components compared to other non-bilateria (Figure 3.4). Furthermore, having examined four species, I was able in part to compensate for absences in single species. In general, there is no clear-cut distinction between rhabdomeric profile

or ciliary profile. *Stylophora pistillata* and *Nematostella vectensis* both express ciliary type opsins, while *Hydra vulgaris* and *Clytia hemisphaerica* express opsins that are RGR/Go type according to my phylogenetic analysis (see supplementary files with the full reconciliation). The opsin expression may suggest a potentially more ciliary-like profile as has been suggested (Plachetzki et al. 2010). However, the overall difficulty in distinguishing between rhabdomeric and ciliary profile may reflect a growing view that cnidaria possess a different pathway, that while sharing some components with the two traditional cascades, also includes cnidaria-specific elements yet to be characterised (Vöcking et al. 2022).

PRC-like in Placozoa

The placozoan *Trichoplax adhaerens* has a very simple body plan in which only a handful of cell types have been described morphologically (Smith et al. 2014), although molecular studies have uncovered a broader diversity of cell types (Sebé-Pedrós, Chomsky, et al. 2018; Varoqueaux et al. 2018). While *Trichoplax* seems to have at least some basic response to light (Heyland et al. 2014), there is no morphological evidence of the presence of photoreceptor cells. Furthermore, while their placopsins originated from the same gene duplication as other animal opsins (Feuda et al. 2012), they do not possess a retinal binding domain, complicating our understanding of whether they can actually serve in light detection. Bearing this in mind, here the goal was to test whether I could at least find any PRC-like profile that could be further explored as candidate homologous cell type to PRCs, whether or not it may indeed have a role in light response. The single cell data analysis (see methods) highlighted 5 candidate metacells (Figure 3.4). Interestingly, from the *Trichoplax* genome I identified representatives of all eleven rhabdomeric gene families (Figure 3.2) and these were all detected in the single cell data except the MYO3/16 family (Figure 3.4). This contrasts with the ciliary genes, of which only a handful were present in the genome. Although this asymmetry complicates the comparison between potential rhabdomeric versus ciliary profiles, it is important to note that most rhabdomeric components are expressed in the *Trichoplax* PRC-like metacells. Further functional exploration of this cascade could therefore be of relevance in the future.

PRC-like in Porifera

Although sponges lack opsins and, like placozoans, do not possess neurons, they are known to be receptive to light (Leys and Degnan 2001; Maldonado et al. 2003; Elliott and Leys 2004; Wong et al. 2022). It has been proposed that sponges may utilise light sensitive cryptochromes (Rivera et al. 2012; Müller et al. 2013), or even other GPCRs such as glutamate receptors (Wong et al. 2022), instead of opsins for photoreception. Furthermore, in *Amphimedon queenslandica* two rhabdomeric phototransduction genes have been implicated in phototactic behaviour of the larvae (Wong et al. 2022), further suggesting the existence of a phototransduction pathway and potentially a photoreceptor cell type in these animals. From the phylogenetic analysis, I found that a couple of rhabdomeric genes and most ciliary genes were missing from the *Amphimedon* genome. Overall, this species, together with the ctenophore (see below), is the one with fewest phototransduction genes recovered in the genome. In the PRC-like metacells that were recovered from the single cell analysis, the few ciliary genes found in the genome are all expressed, as are all the common genes, and most of the rhabdomeric genes. Due to the paucity of ciliary genes in comparison to the rhabdomeric genes, we may be inclined to suggest that a rhabdomeric-like profile is predominant. However, like for cnidaria, it could be that sponges utilise some components of the classic phototransduction cascades alongside more lineage-specific components.

PRC-like in Ctenophora

In the ctenophore *Mnemiopsis leidyi*, a morphologically ciliary-type photoreceptor cell (Horridge 1964; Tamm 2016) has been reported. Although PRCs are not entirely characterised, ctenophores are generally considered to be more likely to possess PRCs compared to placozoans and sponges, as they have neurons and complex behaviours that include predation (Jékely et al. 2015). Importantly, *M. leidyi* has opsins ((Schnitzler et al. 2012; Feuda et al. 2014) and this study) which is another clue that there might be functional PRCs, although we do not know if the phototransduction pathway used might be similar to one of the already described ones or could be independent. Here I find 4 candidate metacells (Figure 3.4). As many phototransduction genes, especially ciliary ones, were missing from the genome, it is difficult to make strong conclusions. Although more rhabdomeric genes were present in the genome compared to ciliary genes, in the PRC-like metacells, the few ciliary genes are almost entirely expressed, in contrast to the

rhabdomeric genes that are expressed in less metacells. The most extreme case is in metacell 39 that expresses all three ciliary genes it has in the genome but only two of the eight rhabdomeric genes available in the genome. A previous study (Schnitzler et al. 2012) reported to have found many ciliary phototransduction genes in *Mnemiopsis leidyi*, in contrast to only a handful of rhabdomeric genes. Overall that study reported more phototransduction genes than the ones reported here, however, their data mining was exclusively based on BLAST, with phylogenetic analysis dedicated only to the opsin gene, so likely some of those genes were filtered out in the more rigorous phylogenetic analysis conducted here. In any case, their conclusion that *M. leidyi* PRCs have a ciliary type phototransduction is compatible with my results, although we must caution that possibly ctenophores have some alternative specific components in their cascade.

Shared regulatory toolkit of PRC-Like metacells throughout animals.

The putative PRCs I identified throughout animals were based on the expression of phototransduction genes. This helped to identify cells that may have the molecular machinery to perform phototransduction and are therefore similar to known PRCs at least from a potentially functional perspective. However, to explore the potential homology amongst cell types across species, we must focus on the core regulatory complex of the cells, namely the set of genes, such as transcription factors, that regulate the expression of other genes and determine the cell identity (Arendt et al. 2016).

I collected a list of orthogroups of regulatory genes that are differentially expressed in each of the PRC-like metacells (see Methods) and used this information to further understand relationships amongst metacells across species.

Orthogroups of regulatory genes

A total of 806 EggNog orthogroups (see Methods) for regulatory genes were identified amongst the highly expressed genes of the PRC-like metacells of the 12 species examined. On average, each species possessed more orthogroups that were shared with at least one other species rather than species-specific ones (Figure 3.5), suggesting some degree of communal regulatory profile amongst PRCs across animals. Orthogroups that

were species-specific were often metacell-specific within that species as well (Figure 3.6). Metacell-specific orthogroups are unlikely to be indicative of a universal core PRC cell profile, therefore, these were discarded from further analyses. This left 421 orthogroups that were shared across at least two metacells (Figure 3.7). Of these, 286 orthogroups were present in at least two species (Extended Figure 3.8A, available on GitHub), 219 were present in at least 2 phyla and 69 are in 3 or more phyla (Figure 3.8A).

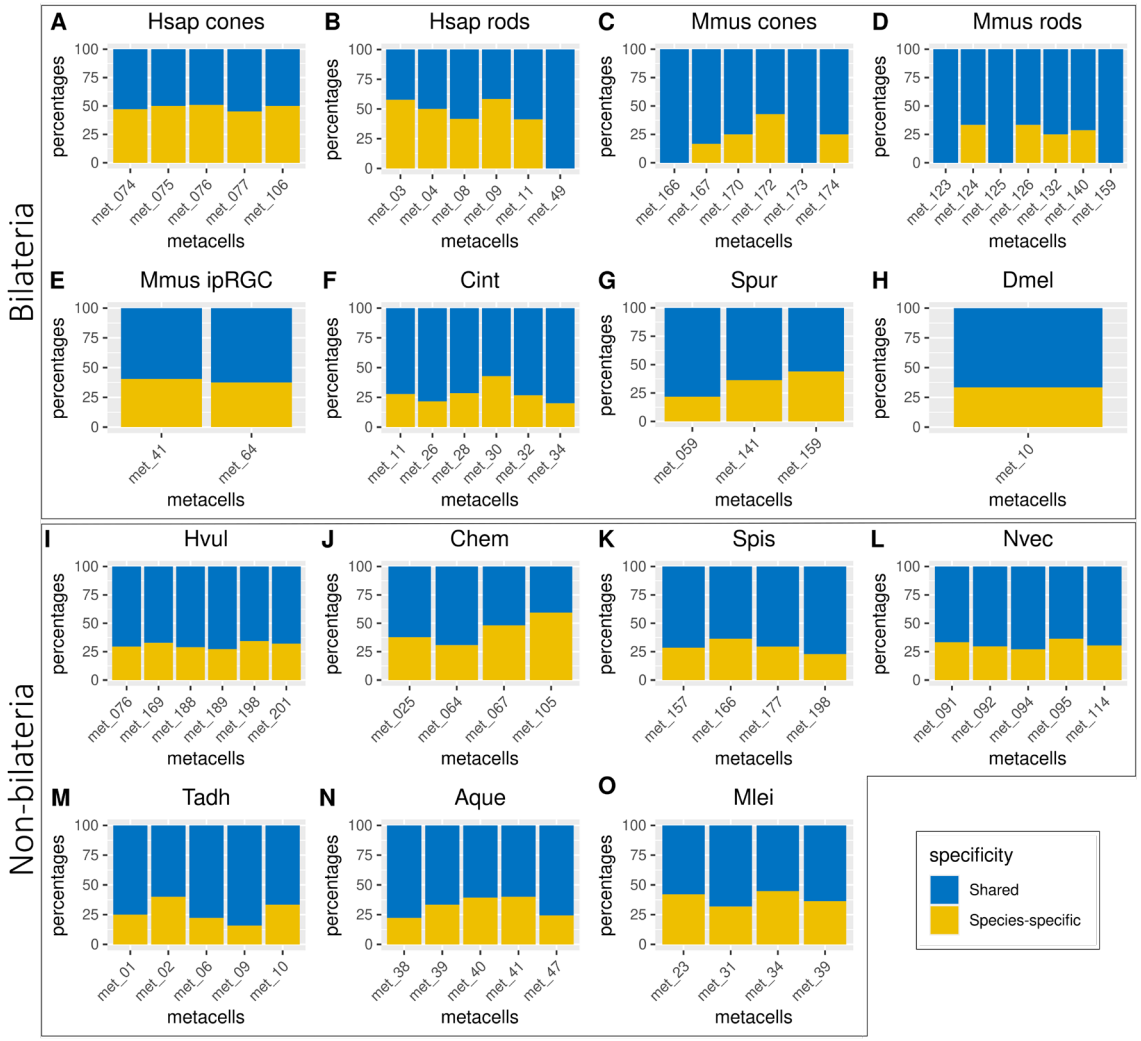


Figure 3.5. Shared versus species-specific orthogroups of regulatory genes in PRC-like metacells throughout animals. For each PRC-like metacell of each species a list of the regulatory genes expressed was compiled. Regulatory genes orthogroups were defined through EggNog (see Methods). Cross-species comparison revealed that the majority of the orthogroups examined are shared with at least one other species. This was true both in bilaterian and in non-bilaterian animals. Abbreviations: *D. mel*: *Drosophila melanogaster*; *H. sap*: *Homo sapiens*; *M. mus*: *Mus musculus*; *C. int*: *Ciona intestinalis*; *S. pur*: *Strongylocentrotus purpuratus*; *N. vec*: *Nematostella vectensis*; *S. pis*: *Stylophora pistillata*; *C. hem*: *Clytia hemisphaerica*; *H. vul*: *Hydra vulgaris*; *T. adh*: *Trichoplax adhaerens*; *A. que*: *Amphimedon queenslandica*; *M. lei*: *Mnemiopsis leidyi*.

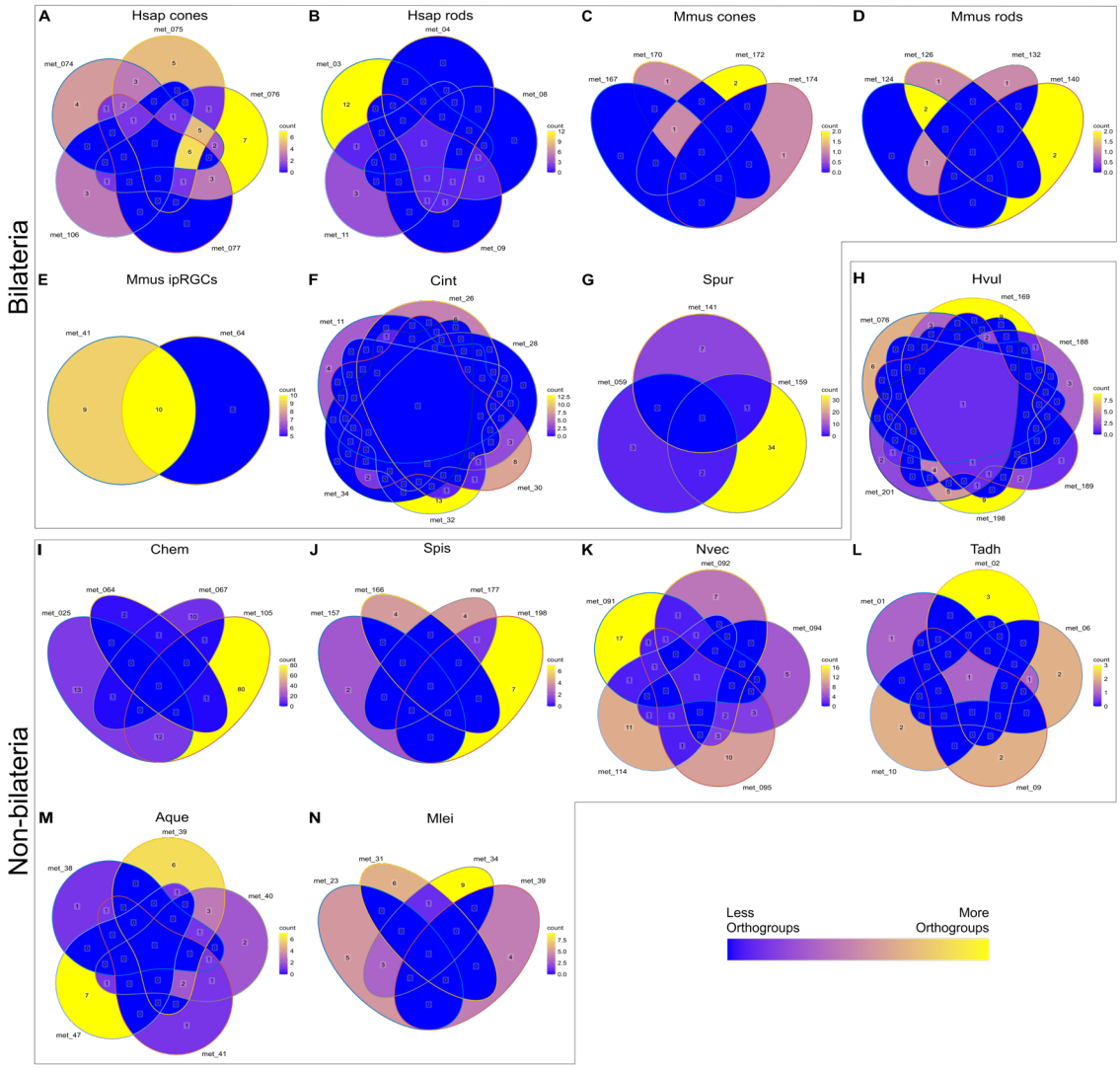


Figure 3.6. Species-specific orthogroups of regulatory genes across PRC-like metacells for each species. Venn diagrams were made to visualise how many orthogroups were shared amongst metacells of the same species. In many cases, species-specific orthogroups of regulatory genes are also metacell-specific or at least shared by only a small subset of PRC-like metacells within a given species. The most striking exception is within the ipRGCs of mouse (**E**), were the majority of the species-specific orthogroups are shared between the two ipRGC metacells of mouse. Human (**A-B**) and mouse (**C-E**) metacells are divided here by PRC type (cones, rods or ipRGCs) to maintain the graphic representation human-readable; in any case orthogroups that are not shared with same type PRCs are generally not shared with different PRC types. The fly (*D. melanogaster*) is missing here, as it has only one PRC-like metacell. Abbreviations: *H. sap*: *Homo sapiens*; *M. mus*: *Mus musculus*; *C. int*: *Ciona intestinalis*; *S. pur*: *Strongylocentrotus purpuratus*; *N. vec*: *Nematostella vectensis*; *S. pis*: *Stylophora pistillata*; *C. hem*: *Clytia hemisphaerica*; *H. vul*: *Hydra vulgaris*; *T. adh*: *Trichoplax adhaerens*; *A. que*: *Amphimedon queenslandica*; *M. lei*: *Mnemiopsis leidyi*.

Structure of relationships amongst PRC-like metacells

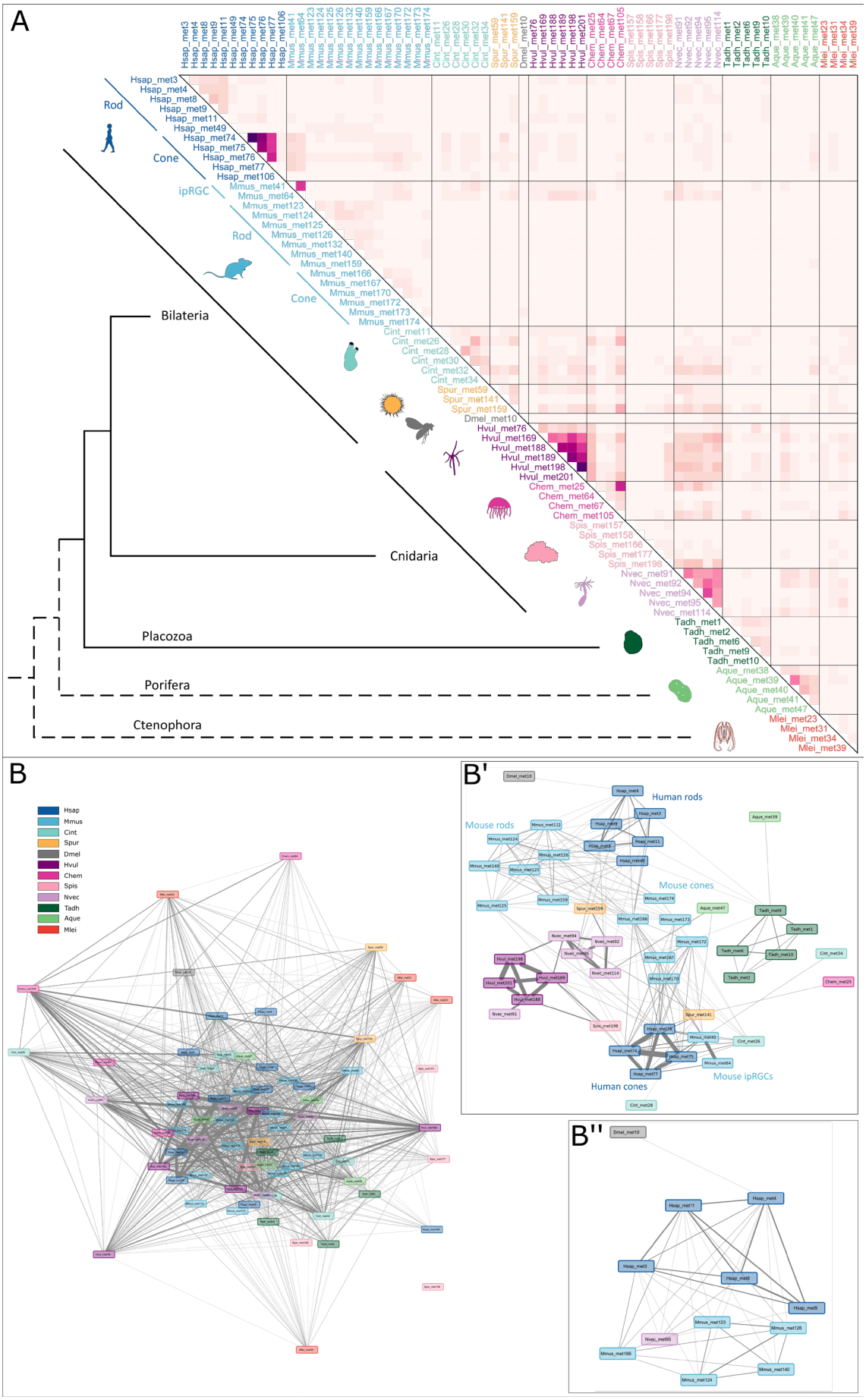
As a first step in comparing the regulatory toolkit of animal PRC-like metacells, I investigated how many orthogroups were shared and by which metacells (Figure 3.7A). Whilst the highest number of shared genes is between PRCs of the same or closely related species, there are still several shared genes also amongst distantly related species. To have a broader understanding of the relationships amongst PRCs of different species, I constructed a network to visualise connections amongst metacells based on the number of shared regulatory genes (Figure 3.7B). This network was built using only the regulatory genes that were amongst the top 100 highly expressed genes for each metacells, as this would provide higher confidence connections (see Methods). The strength of the network approach is both to obtain an overview of metacell relationships and to identify indirect connections that are otherwise difficult to spot. The network of all

metcells revealed a vast number of connections linking metacells either directly or indirectly. To better discern the relationships amongst a subset of metacells, I extracted subnetworks of the metacells most closely related to human PRC metacells (Figure 3.7B') and to the *Drosophila melanogaster* PRC metacell (Figure 3.7B'').

From the human PRCs subnetwork (Figure 3.7B'), we can observe that human rods cluster together and are closely connected to mouse rods. Similarly, human cones are strongly clustered together and connect to mouse cones. While the connection between mouse rods and cones is solid, the direct connection between human rods and cones is weaker. Three *Ciona intestinalis* metacells are directly connected to various combination of metacells of the broad cluster containing human and mouse cones as well as mouse ipsRGC and a sea urchin metacell. Curiously, these urochordate metacells are not directly connected to each other. Two sea urchin metacells are also quite related to human and mouse PRCs, one to the cone type and the other to the rod type. A cluster of cnidarian metacells appears connected to both rod and cone clusters. A *Trichoplax adhaerens* cluster has a few connections with the rod cluster. The sponge metacell 39 is loosely connected to rod type PRCs via the *Trichoplax* cluster. Curiously, the *Drosophila* metacell has a connection with one human metacell of the rods cluster but no connection to the two mouse ipRGCs (41 and 64), that are candidate homologs to rhabdomeric PRCs, that instead cluster more closely to the cones cluster. Therefore, while these rhabdomeric-like mouse PRCs may utilise a rhabdomeric-like cascade, from a regulatory perspective they do not appear to share a similar identity to the classic rhabdomeric cell type of *Drosophila*.

Drosophila melanogaster metacell 10 is the only representative of the rhabdomeric type PRC. The subnetwork of this metacell with its closest related metacells (Figure 3.7B'') confirms the connection to the rod ciliary cluster through direct connection to only one human metacell. No other relationship with the rest of the dataset is detected. Whilst this suggests a unique transcription factor profile for this PRC type, it is important to note that this network analysis was conducted considering only the top 100 highest expressed genes of each metacell, therefore, connections based on shared regulatory genes with lower expression do not appear.

Figure 3.7. Cross-species comparison of orthogroups of regulatory genes expressed in PRC-like metacells. Orthogroups of regulatory genes were identified with EggNog. (**A**) Heatmap showing the number of orthogroups in common amongst PRC-like metacells across species. While the majority of the shared orthogroups are amongst metacells of the same or closely related species, there is still some degree of shared orthogroups amongst distantly related species. (**B**) A network analysis of the orthogroups in common among metacells. Nodes are metacells and edges represent shared regulatory genes. The thickness of the edges is proportional to the number of shared genes. This analysis highlighted many indirect connections, indicating some level of relationship across all PRC-like metacells. (**B'**) A subnetwork of human PRC metacells and their most closely related metacells (first two neighbours) reveals details about the relationships between potentially ciliary type metacells. Of note, the mouse ipRGCs (candidate rhabdomeric PRCs) appear more similar to human cone PRCs rather than to the *Drosophila* rhabdomeric metacell. (**B''**) A subnetwork with the *Drosophila* PRC and its closest relatives (first two neighbours) does


A
Rod
Cone
ipRGC
Rod
Cone
Bilateria
Cnidaria
Placozoa
Porifera
Ctenophora
B
Hsap
Mmus
Cint
Spur
Dmel
Hvul
Chem
Spis
Nvec
Tadh
Aque
Mlei
B'
Human rods
Mouse rods
Mouse cones
Human cones
Mouse ipRGCs
B''

Species-specific combinations of regulatory genes across Metazoan PRC-like metacells

Next, I examined which genes were responsible for the above network connections. Interestingly, from a vast list of hundreds of orthogroups of regulatory genes, only 69 were expressed in 3 or more phyla (Figure 3.8A), while the majority were expressed in 2 or 1 phyla (Extended Figure 3.8A). Of the regulatory genes in common between 3 or more phyla, some, like Six6/3, Meis2 and Tbx2 for ciliary type (Zuber et al. 2003; Alvarez-Delfin et al. 2009; Vopalensky and Kozmik 2009) and glass for rhabdomeric (Bernardo-Garcia et al. 2017), are recognised for their role in PRC identity and/or specification. However, for others there is no known connection. Furthermore, some transcription factors that are well known to be involved in photoreceptor identity/specification, for example Otx or Rx (Arendt 2003; Vopalensky and Kozmik 2009), did not pass the threshold of 3 or more phyla in my dataset (Extended Figure 3.8A). Curiously, while there seems to be some conserved pattern of combinations of regulatory genes expressed within PRC-like metacells of the same species, across different species there seems to be little conservation. This explains all the indirect connections found in the network. Ultimately all metacells are "related" to each other to some degree because they share one or few genes with a metacell that in turn shares another set of few genes with a different metacell and so on. The results of this comparative analysis suggest that the core regulatory complex of PRC-like metacells in different species comprises a set of species-specific genes. Although some regulatory genes make a recurrent presence across species, the exact combination of genes is often different.

Transcription factors amongst the most abundant regulatory genes in PRC-like metacells throughout animals

Finally, I classified the orthogroups of regulatory genes into finer categories by performing BLASTP versus the Animal Transcription Factor Database (ATFDB v4) (Shen et al. 2023). This clarified that amongst the 69 orthogroups that are shared across 3 or more phyla, around 60% belonged to transcription factor families (Figure 3.8B). The remainder of the orthogroups were either transcription cofactors (~23%), genes that interact with transcription factors but do not directly bind DNA, or other regulatory genes (~17%) like RNA polymerases and proteins that interact with the chromatin structure. The transcription factors belonged to multiple different families, with bZIP transcription

factors, zinc finger C2H2 and homeoboxes being amongst the most predominant (Figure 3.8C). Overall, the most abundant transcription factor families possess varied types of DNA binding domains, but the most popular are the basic domains and the helix-turn-helix, although many others have an unclassified structure. Examining all 421 orthogroups that were present in at least 2 metacells (Supplementary Table S3.3), the percentages of these categories were very similar: ~59% transcription factors; ~29% cofactors; ~12% other regulatory genes. With zinc finger C2H2 and homeoboxes being again amongst the most frequent transcription factor families. bZIP were still very abundant, although marginally surpassed by HLH and bHLH.

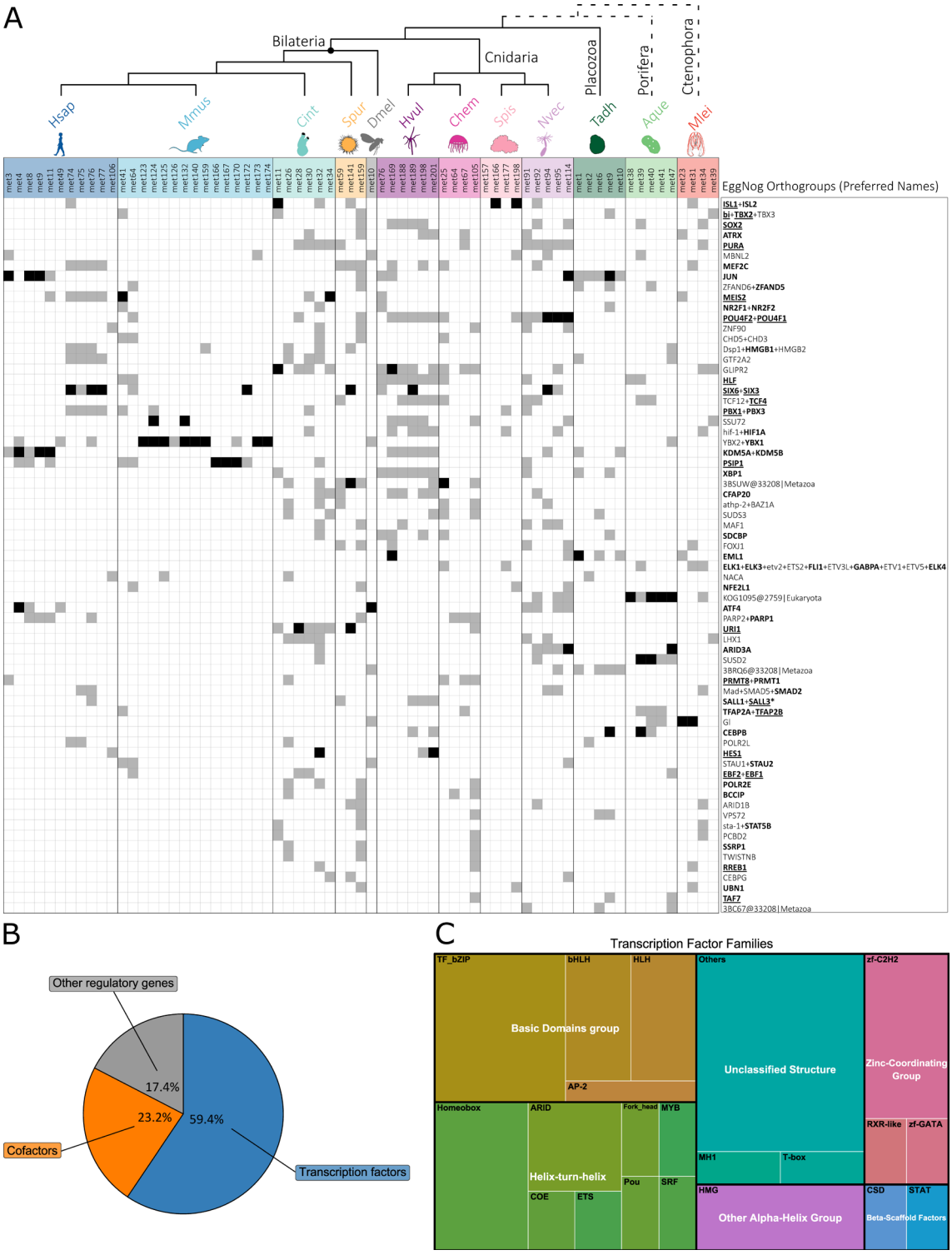


Figure 3.8. Most common orthogroups of regulatory genes shared across PRC-like metacells throughout animals. (**A**) 69 orthogroups are present in 3 or more phyla. Orthogroups are ordered by most frequent (with the hierarchy: present in most phyla, present in most species, present in most metacells). While some orthogroups are frequently expressed throughout animals, their exact combination of co-expression varies in the different species. Genes that were amongst the top 100 differentially expressed genes of a metacell are indicated with a black square; a grey square indicates expression with log-fold change (lfp) > 0.5. Therefore, black squares indicate strong markers for a given metacell, while grey squares indicate that the gene is expressed in the metacell but differential expression level is not as high. Names derive from the Preferred_names of the respective EggNog orthogroups, where present, or the EggNog orthogroup itself. GeneCards/Flybase were used to characterise the human/*Drosophila*

representatives. Genes are highlighted: in bold if there is some evidence of involvement in vision and/or eye/photoreceptor development; in bold and underlined if there is strong evidence of involvement in vision and/or are expressed in the retina, although not necessarily in cones and rods; in bold, underlined and with asterisk if they are specifically expressed in photoreceptor cells. Species silhouettes were modified from images with CC0 1.0 Universal Public Domain Dedication licences obtained from https://www.phylopic.org/. Abbreviations: *D. mel*: *Drosophila melanogaster*; *H. sap*: *Homo sapiens*; *M. mus*: *Mus musculus*; *C. int*: *Ciona intestinalis*; *S. pur*: *Strongylocentrotus purpuratus*; *N. vec*: *Nematostella vectensis*; *S. pis*: *Stylophora pistillata*; *C. hem*: *Clytia hemisphaerica*; *H. vul*: *Hydra vulgaris*; *T. adh*: *Trichoplax adhaerens*; *A. que*: *Amphimedon queenslandica*; *M. lei*: *Mnemiopsis leidyi*. (**B**) The majority of the orthogroups of regulatory genes are transcription factor families (59.4%). Transcription cofactors are also abundant (23.2%). The remaining orthogroups include a mixture of other genes that are involved in transcription, such as polymerases and genes involved in chromatin conformation. (**C**) Treemap of the most abundant families of transcription factors shared across PRC-like metacells, organised by broad groups based on the type of DNA-binding domain. Basic domains and helix-turn-helix are the most abundant.

Conclusions

The comprehensive analysis of the evolution of phototransduction genes revealed that their broad families mostly originated anciently in eukaryotes. Notably, even the sub-lineages containing the precise genes specialised in phototransduction functions often trace back to pre-Metazoan/Holozoan times, with few exceptions primarily amongst ciliary components. This in turn has important implications for understanding the evolution of the photoreceptor cell type in which phototransduction is employed.

Using the phototransduction genes found in non-model organisms, including all non-bilaterian phyla, I was able to detect photoreceptor cell-like profiles in their single-cell dataset. In early branching animals a mixed situation in the expression of the core components of either one or both the classical rhabdomeric/ciliary pathways, suggests that some shared components were likely employed early on in phototransduction, but then different animal lineages recruited a specific set of other components. Future research should therefore focus on uncovering these species-specific phototransduction variants in early branching animals (Vöcking et al. 2022), as well as in a more diverse set of bilaterian species.

Furthermore, the analysis of regulatory genes differentially expressed in these photoreceptor-like cells uncovered that the most common category of regulatory genes shared across animal PRCs are transcription factors. The results suggest that the exact combinations of these genes are species-specific. Nevertheless, the recurrence of some transcription factors associated to PRCs of model organisms, in some, albeit not consistently all, non-bilaterian PRC metacells, e.g., SOX and TBX homologs, supports the notion that some core transcription factors might have contributed to the identity of an ancestral PRC type. Moreover, there are some transcription factor families that are more abundant than others. For example, bZIP transcription factors, zinc fingers C2H2 and homeoboxes are amongst the most frequent. These transcription factor families in turn deploy varied DNA-binding-domains, suggesting a broad spectrum of mechanisms through which transcription can be regulated in photoreceptor-like cells throughout animals.

Finally, this work compiled an extensive list of molecular components that could be involved in phototransduction and photoreceptor-like cell identity in non-bilaterians. This can be used as a valuable resource in future research on the functional characterisation of visual systems in these organisms.

Methods

Reconstruction of the Evolution of Phototransduction Components.

Species List and Species Tree

To investigate the deep origin of the gene families of the phototransduction components, the search was broadened to all eukaryotes. 86 species representatives of Eukarya were chosen based on proteome completeness and taxonomic sampling (Table 3.2). Focus was given to sister taxa of Metazoa (8 choanoflagellates and 5 other holozoans) and non-bilaterian Metazoa (25 species), since functional visual processes must have originated at an early stage of animal evolution. The proteome completeness was assessed with BUSCO (v4.0.6) (Simão et al. 2015; Waterhouse et al. 2018) using the eukaryota_odb10 database of 255 BUSCO genes (See Supplementary Table S3.1 with proteome details).

Prior knowledge of species relationships can provide a backbone for species-tree-aware gene tree construction (Boussau and Scornavacca 2020). Therefore, I used the BUSCO genes from each species for the construction of a species tree. Briefly, BUSCO genes were extracted and aligned with MAFFT v7.470 (--auto) (Katoh et al. 2002; Katoh and Standley 2013) and trimmed with Trimal v1.4.rev22 (-automated1) (Capella-Gutiérrez et al. 2009). Trimmed alignments of all BUSCO genes were concatenated with FASconCAT v1.11 (Kück and Meusemann 2010) into a super-matrix. The super-matrix was used as input for species tree construction with IQTREE v2.0.6 (Hoang et al. 2018; Minh et al. 2020), after running Model Finder (Kalyaanamoorthy et al. 2017) for best-fitting model. The resulting species tree was inspected to confirm that known species and phyla relationships were recovered. The species tree obtained places Ctenophores as sister group to all other metazoans. As this is one of the currently accepted scenarios (Whelan et al. 2017; Schultz et al. 2023), this topology was kept. The alternative topology (Sponges as sister-group to all other animals) (Feuda et al. 2017) was obtained by manually swapping branches with Mesquite v3.6.1 (Maddison and Maddison 2008). Both species topologies were kept for downstream applications (and are available on GitHub).

Data Mining

Molecular components of interest were based on *Drosophila melanogaster* and *Homo sapiens* pathways as representative of rhabdomeric and ciliary phototransduction respectively. Some elements of the pathways are composed of multiple subunits encoded by different genes. In total 28 gene families were identified based primarily on the KEGG maps ko04745 (rhabdomeric) and ko04744 (ciliary) (Kanehisa et al. 2021). Two additional genes, the RGS9BP and GNB5 subunits of the RGS9 complex, were added based on updated references of vertebrate phototransduction (Lamb et al. 2018) (Figure 3.1 and Table 3.1). Queries were collected from the KEGG Orthology lists (Kanehisa 2019) for each component present in the KEGG pathways and from (Lamb et al. 2018) for the two additional gene families. BLASTP (Camacho et al. 2009) was conducted (e-value cut-off of 1e-5) for each query versus the species database. Potential duplicates were removed with cd-hit (Li et al. 2001; Fu et al. 2012) with identity threshold of 100%. Outputs were used for another BLASTP versus the SwissProt database (Poux et al. 2017). Sequences were kept only if the gene family of interest was within the top five hits and parsing was carried out with gene family-specific keywords (See Supplementary Table S3.4 with list of keywords per component). This provided a first level of similarity-based filtering. A second round of filtering was conducted based on the presence of gene family-specific protein motifs. The filtered dataset was scanned with InterProScan (Quevillon et al. 2005; Jones et al. 2014) and sequences were kept only if they contained the combination of motifs characteristic to their gene family (See Supplementary Table S3.4 with list of protein motifs). To provide an annotation to the final collections of sequences, I used the top hit from BLASTP versus SwissProt.

Phylogenetic Trees

Gene trees constructed for each gene family followed a standard pipeline: alignment of sequences with MAFFT (--auto) (Katoh et al. 2002; Katoh and Standley 2013); trimming of sequences to eliminate columns with more than 70% gaps (Trimal with -gt 0.3) (Capella-Gutiérrez et al. 2009); tree construction after running Model Finder in IQTREE2 (Kalyaanamoorthy et al. 2017). The list of models tested and the best-fit models for each family, chosen based on the Bayesian Information Criterion (Schwarz 1978), can be found in Supplementary Table 3.5. To obtain fully bifurcating gene trees necessary as inputs for

the gene tree to species tree reconciliations (see below), any polytomy in the gene trees was randomly resolved with ETE3 (Huerta-Cepas et al. 2016).

Gene tree to species tree reconciliation

The resulting gene trees were used as starting trees for a gene tree to species tree reconciliation using Generax (v1.2.3) (Morel et al. 2020). The model used to compute the reconciliation was set to account for duplication and loss, but not transfer events. Both alternative species trees (ctenophore-first and sponge-first) were tested. The number of duplications and losses were extracted for each gene family and compared between ctenophore-first and sponge-first scenarios (see Supplementary Table S3.2).

Resulting reconciled trees were manually examined to trace the evolution of the genes of interest. The *D. melanogaster* and *H. sapiens* genes known to function in phototransduction were used to identify the orthogroups of interest and the duplication and loss events that characterised their lineages. Other subgroups within the gene families and their relationship with the orthogroups of interest were also identified. Comparison between the two alternative reconciliations with ctenophore-first versus sponge-first species tree provided a more comprehensive picture for the reconstruction of the evolutionary history of the gene families.

Collection of phototransduction marker genes for photoreceptor cells in non-model organisms

By tracing the presence of the orthogroups of interest, as identified with the reconciliations, throughout all the species examined, I was able to collect a list of candidate marker genes for phototransduction also in non-model organisms. Where the orthogroup of interest was not present, closely related lineages were used as potential markers. These marker genes were used for identifying candidate photoreceptor cell types in non-model organisms, including several non-bilaterians, for which single-cell RNA sequencing data was available. See more details below.

Identification of putative photoreceptor cell types from single-cell RNA-sequencing data.

Species datasets

To obtain a sample of photoreceptor cell diversity throughout Metazoa, I focused the single-cell analysis on twelve species based on scRNAseq data availability and phylogenetic representation. *Drosophila melanogaster* (Özel et al. 2021) served as an example for rhabdomeric-type PRCs, while *Homo sapiens* (Lukowski et al. 2019) and *Mus musculus* (Macosko et al. 2015) were representative for ciliary-type PRCs. Two additional deuterostomes (the urochordate *Ciona intestinalis* (Sharma et al. 2019) and the sea urchin *Strongylocentrotus purpuratus* (Paganos et al. 2021) served as bridge species between vertebrate PRCs and protostome PRCs as represented by *Drosophila*. Finally, of particular interest for this project are non-bilaterian animals: I therefore included four cnidarian species (*Hydra vulgaris* (Siebert et al. 2019), *Clytia hemisphaerica* (Chari et al. 2021), *Stylophora pistillata* (Levy et al. 2021) and *Nematostella vectensis* (Sebé-Pedrós, Saudemont, et al. 2018)), the placozoan *Trichoplax adhaerens* (Sebé-Pedrós, Chomsky, et al. 2018), the sponge *Amphimedon queenslandica* (Sebé-Pedrós, Chomsky, et al. 2018), and the ctenophore *Mnemiopsis leidyi* (Sebé-Pedrós, Chomsky, et al. 2018). The details of these scRNAseq datasets are summarised in Table 3.3.

MetaCell pipeline for clustering cells

For the search of photoreceptor-like cells in the species of interest, I used the approach of identifying "metacells" or cell states to account for potential low depth of sequencing in non-model organisms, especially when the dataset is of the whole body.

Unique Molecular Identifiers (UMI) count matrices for each species were used as input for an established pipeline using the MetaCell v0.3.6 (Baran et al. 2019) R package, as described on MetaCell GitHub (Analyzing whole-organism scRNA-seq data with metacell • metacell (tanaylab.github.io). Once the metacells were computed, heatmaps for all the species-specific phototransduction markers were generated to visualise which metacells were overexpressing them and indeed whether they were co-expressed in the same metacell. To better visualise the situation for single genes, I also generated bar plots with the log fold change values (lfp) of each gene in each metacell and 2D graphs with the expression of single genes mapped into the metacells 2D graph (see supplementary

figures on GitHub). Finally, complete lists of lfp values for all genes in all metacells for each species were extracted for downstream analysis. See metacell scripts on GitHub for each species to reproduce these gene lists as well as the figures.

Identification of photoreceptor metacells in the model organisms *D. melanogaster*, *H. sapiens* and *M. musculus*

As a first step, I tested the pipeline on model organisms to determine whether photoreceptor cells (PRCs) could reliably be identified. *D. melangoster* rhabdomeric phototransduction genes were used to pinpoint a rhabdomeric PRC profile; and ciliary phototransduction genes of *H. sapiens* and *M. musculus* were used to identify ciliary-type PRCs. In the case of human and mouse, since it has been proposed that OPN4 (melanopsin) expressing cells, such as retinal ganglion cells, of vertebrates are homologous to rhabdomeric PRCs (Provencio et al. 2000; Hattar et al. 2002; Arendt 2003; Rollag et al. 2003; Fu et al. 2005), I searched also for candidate rhabdomeric PRC profiles. For this I used OPN4 (an r-opsin) together with the other rhabdomeric genes that were found in human and mouse as markers.

In the case of *D. melanogaster*, the identification of a rhabdomeric PRC profile was extremely straightforward. It was possible to spot a candidate metacell already with the heatmap of phototransduction marker expression. This metacell was kept as an example for rhabdomeric PRC-type for comparison with non-model organisms (see below). Conversely, in human and mouse datasets, multiple metacells were good candidate PRCs. This was likely due to the fact that both datasets used were from retinal samples and it is indeed expected to identify multiple PRC profiles, especially rods that are known to be more abundant than cones. Instead, the *Drosophila* dataset came from an entire optic lobe, where we do expect a more diverse set of cell types. Although it is sensible to keep in consideration several metacells per species as PRC candidates (as effectively each metacell is a cell state so there could be several PRC cell states in the dataset), it is still necessary to discriminate between PRC cells and non-PRC cells present in the retina. Therefore, further steps to decide which metacells to keep were carried out for human and mouse. In order to be consistent with the non-model organisms, the same pipeline was used and is described below.

Identification of candidate photoreceptor metacells in non-model organisms

As identifying photoreceptors in non-model organisms is not straightforward, particularly for some non-bilaterians for which we do not even have any evidence that there may be photoreceptors at all (e.g. placozoa (Smith et al. 2014)), I developed a pipeline to pick-up the metacells that could be most likely a PRC-type. The objective was to determine for which metacells there is sufficient evidence, based on the expression of combinations of phototransduction genes, to say that they have a PRC-like profile.

First, I filtered out metacells in which opsin log fold change (lfp) was below 0.2. This is because the opsin is the strongest marker for a photoreceptor cell, so we expect it to be at least slightly overexpressed. The exception was *Amphimedon queenslandica*, as sponges do not possess opsins ((Feuda et al. 2012) and this chapter). To detect potential photoreceptor cell homologs in sponge the other phototransduction genes were used as only markers. I also ranked all metacells based on highest differential expression (lfp) of an opsin.

The next step was to assess the level of phototransduction gene expression in the metacells. For this I checked both the percentage of phototransduction genes co-expressed in the same metacell and their level of differential expression within the metacell. Specifically, I calculated the percentage of phototransduction genes expressed and their average lfp for: all genes; all common genes; all rhabdomeric genes; and all ciliary genes. Between the latter two, I kept the highest value assuming that metacells lean more towards either a rhabdomeric or a ciliary profile.

Therefore, to classify metacells into best PRC candidates all of the following evidences were available: 1) lfp of highest expressed opsin in metacell; 2) average lfp of all phototransduction genes; 3) average lfp of common phototransduction genes; 4) average lfp of either ciliary or rhabdomeric genes (whichever is highest); 5) highest percentage of all phototransduction genes; 6) highest percentage of common phototransduction genes; 7) highest percentage of either ciliary or rhabdomeric phototransduction genes (whichever is highest).

For each of these categories of evidence, I ranked the metacells from best (1st) to worst (nth). The ranking values for all the metacells were then summed to obtain a final ranking. For all rankings, if metacells tied, they got the same ranking value. I decided to keep as best candidate PRCs to be used for further analyses the PRCs that are in the top 5 of the

final ranking. As a result, circa 5 metacells were retained for each species. Some species have less because less than 5 metacells passed the initial threshold of opsin >0.2. Other species have slightly more than 5 metacells because some metacells tied in the final ranking. Supplementary Table S3.6 shows these ranking calculations, and also an alternative ranking system. In the latter case, metacells were ranked based on how many times they appeared in the top 5 of each of the separate categories. In most cases final best metacells correspond between the two methods. The alternative method is shown for completeness.

Note that in the case of mouse and human ciliary PRCs, this procedure was done separately for rod and cone metacells and the top 5 were collected for both types as indeed their genetic profile can be a bit different and in this way we have full representation of the ciliary type.

Exploration of the regulatory genes' toolkit of candidate PRCs and comparison across species.

After having identified PRC-like metacells based on the expression of phototransduction genes as markers (see previous sections), I then moved on to further characterise the genetic profile of these candidate PRCs. The focus of the analysis was on regulatory genes, such as transcription factors, as these genes influence the rest of the genetic profile of the cell and are considered the core regulatory complex that defines cell identity (Arendt et al. 2016).

For all candidate PRCs of all species, I collected: i) the top 100 most highly expressed genes, these should be considered as additional markers for the metacell; and ii) all genes that have lfp above 0.5, these represent genes that are mildly overexpressed in the given metacell.

Identifying regulatory genes in PRC-like metacells

To identify genes involved in transcription, I used two tools. First, I annotated all the collected genes with Eggnog mapper (Cantalapiedra et al. 2021). Only the genes that fell into the COG category K were kept, as that indicates that they are involved in regulating transcription. Contemporarily, the sequences were scanned for Pfam profiles of known

transcription factors (see Supplementary Table S3.7 with list of profiles searched). These two approaches are complementary, as the first selects genes based on whether they may have a regulatory role in transcription, and the second focuses on collecting genes based on the presence of protein domains known to be present in certain transcription factors. Combining these two approaches, I collected a list of transcription factors and genes involved in transcription for all metacells.

For comparison across species, I used the Eggnog Orthogroup (Eggnog_OG) of the genes. As the comparison is amongst distantly related animals, I chose to compare preferably the Metazoa level of the Eggnog_OG, and only when the Eggnog_OG did not reach Metazoa level, did I collect the most stringent level available (often either Eukarya or Opisthokonta).

Cross species comparison of PRC-like metacells based on shared regulatory genes

To understand the extent to which regulatory genes were shared across species I first made an all-against-all comparison (Figure 3.7A) with all metacells of all species. This was done using genes collected with the lfp cut-off of 0.5.

To better visualise the relationships amongst metacells and species based on shared regulatory genes, I created a network graph using Cytoscape v3.9.1 (Shannon et al. 2003) (Figure 3.7B). To avoid obtaining an over-complicated cluster and to focus on the highest confidence connections, the network was constructed using only the genes that were amongst the top 100 highly expressed for each metacell. (Figure 3.7B). As the network of all metacells from all species still contained too many connections to easily focus on relationships amongst specific subsets of metacells, I extracted subsets of the networks to identify more meaningful connections. So, I extracted the subnetwork containing Human PRCs and the first two neighbouring metacells (directly connecting metacells, and metacells connecting to the directly connected metacells) to explore connections amongst candidate ciliary PRCs (Figure 3.7B'). For candidate rhabdomeric PRCs, I made a subnetwork with *Drosophila* metacell and its next two neighbours (Figure 3.7B'').

Uncovering what type of regulatory genes are most common in PRC-like metacells

The network graphs provided broad information about how many connections are shared amongst metacells, however, I also wanted to understand which genes were behind the connections. For that I mapped the presence/absence of all regulatory genes orthogroups across all species and ordered them by most frequent. (Figure 3.8A and Extended Figure 3.8A).

Furthermore, to understand how many of these orthogroups were transcription factors as opposed to other regulatory genes (e.g., transcription cofactors), I performed a BLASTP versus the Animal Transcription Factor Database (ATFDB version 4) (Shen et al. 2023) database collecting first hits (Figure 3.8B). For the transcription factor orthogroups, I also used this database to categorise them into transcription factor families, in turn distributed across broader groups based on the DNA-binding-domain (Figure 3.8C).

Data Availability

Additional supplementary material and raw output files are available at the GitHub repository: https://github.com/AAleotti/PhD_Thesis.

Acknowledgements

I am extremely grateful to Maryam Ghaffari Saadat and Julien Devilliers for their help in coding and automatising steps used in the methodologies described in this Chapter.

References

Altimimi HF, Schnetkamp PPM. 2007. Na+/Ca2+-K+ Exchangers (NCKX):Functional Properties and Physiological Roles. *Channels* [Internet] 1:62–69. Available from: https://doi.org/10.4161/chan.4366

Alvarez-Delfin K, Morris AC, Snelson CD, Gamse JT, Gupta T, Marlow FL, Mullins MC, Burgess HA, Granato M, Fadool JM. 2009. Tbx2b is required for ultraviolet photoreceptor cell specification during zebrafish retinal development. *Proc. Natl. Acad. Sci.* [Internet] 106:2023–2028. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0809439106

Arendt D. 2003. Evolution of eyes and photoreceptor cell types. *Int. J. Dev. Biol.* 47:563–571.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* [Internet] 9:868–882. Available from: https://www.nature.com/articles/nrg2416

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD, et al. 2016. The origin and evolution of cell types. *Nat. Rev. Genet.* 17:744–757.

Arendt D, Tessmar-Raible K, Snyman H, Dorresteijn AW, Wittbrodt J. 2004. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science* 306:869–871.

Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, Meir Z, Hoichman M, Lifshitz A, Tanay A. 2019. MetaCell: analysis of single-cell RNA-

seq data using K-nn graph partitions. *Genome Biol.* [Internet] 20:206. Available from: https://doi.org/10.1186/s13059-019-1812-2

Bernardo-Garcia FJ, Humberg T-H, Fritsch C, Sprecher SG. 2017. Successive requirement of Glass and Hazy for photoreceptor specification and maintenance in Drosophila. *Fly (Austin)* [Internet] 11:112–120. Available from: https://doi.org/10.1080/19336934.2016.1244591

Boussau B, Scornavacca C. 2020. Reconciling Gene trees with Species Trees. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.2:1-3.2:23. Available from: https://hal.science/hal-02535529

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* [Internet] 10:421. Available from: https://doi.org/10.1186/1471-2105-10-421

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Available from: https://www.biorxiv.org/content/10.1101/2021.06.03.446934v2

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* [Internet] 25:1972–1973. Available from: https://doi.org/10.1093/bioinformatics/btp348

Chari T, Weissbourd B, Gehring J, Ferraioli A, Leclère L, Herl M, Gao F, Chevalier S, Copley RR, Houliston E, et al. 2021. Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across Clytia medusa cell types. *Sci. Adv.* [Internet] 7:eabh1683. Available from: https://doi.org/10.1126/sciadv.abh1683

D'Aniello S, Delroisse J, Valero-Gracia A, Lowe EK, Byrne M, Cannon JT, Halanych KM, Elphick MR, Mallefet J, Kaul-Strehlow S, et al. 2015. Opsin evolution in the Ambulacraria. *Mar. Genomics* [Internet] 24:177–183. Available from: https://www.sciencedirect.com/science/article/pii/S1874778715300349

von Döhren J, Bartolomaeus T. 2018. Unexpected ultrastructure of an eye in Spiralia: the larval ocelli of Procephalothrix oestrymnicus (Nemertea). *Zoomorphology* [Internet] 137:241–248. Available from: https://doi.org/10.1007/s00435-017-0394-3

Eakin RM, Kuda A. 1970. Ultrastructure of sensory receptors in ascidian tadpoles. *Z. Für Zellforsch. Mikrosk. Anat.* [Internet] 112:287–312. Available from: https://doi.org/10.1007/BF02584045

Elliott GRD, Leys SP. 2004. SPONGE LARVAL PHOTOTAXIS: A COMPARATIVE STUDY. *BMIB - Boll. Dei Musei E Degli Ist. Biol.* [Internet] 68. Available from: https://riviste.unige.it/index.php/BMIB/article/view/625

Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Curr. Biol.* [Internet] 27:3864-3870.e4. Available from: https://www.sciencedirect.com/science/article/pii/S0960982217314537

Feuda R, Hamilton SC, McInerney JO, Pisani D. 2012. Metazoan opsin evolution reveals a simple route to animal vision. *Proc. Natl. Acad. Sci.* [Internet] 109:18868–18872. Available from: https://www.pnas.org/content/109/46/18868

Feuda R, Rota-Stabelli O, Oakley TH, Pisani D. 2014. The Comb Jelly Opsins and the Origins of Animal Phototransduction. *Genome Biol. Evol.* [Internet] 6:1964–1971. Available from: https://doi.org/10.1093/gbe/evu154

Fleming JF, Feuda R, Roberts NW, Pisani D. 2020. A Novel Approach to Investigate the Effect of Tree Reconstruction Artifacts in Single-Gene Analysis Clarifies Opsin Evolution in Nonbilaterian Metazoans. *Genome Biol. Evol.* [Internet] 12:3906–3916. Available from: https://doi.org/10.1093/gbe/evaa015

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* [Internet] 28:3150–3152. Available from: https://doi.org/10.1093/bioinformatics/bts565

Fu Y, Liao H-W, Do MTH, Yau K-W. 2005. Non-image-forming ocular photoreception in vertebrates. *Curr. Opin. Neurobiol.* [Internet] 15:415–422. Available from: https://www.sciencedirect.com/science/article/pii/S0959438805001042

Gornik SG, Bergheim BG, Morel B, Stamatakis A, Foulkes NS, Guse A. 2021. Photoreceptor Diversification Accompanies the Evolution of Anthozoa. *Mol. Biol. Evol.* [Internet] 38:1744–1760. Available from: https://doi.org/10.1093/molbev/msaa304

Gurevich VV, Gurevich EV. 2016. G Protein-Coupled Receptor Kinases (GRKs) History: Evolution and Discovery. In: Gurevich VV, Gurevich EV, Tesmer JJG, editors. G Protein-Coupled Receptor Kinases. New York, NY: Springer New York. p. 3–22. Available from: https://doi.org/10.1007/978-1-4939-3798-1_1

Hahn J, Monavarfeshani A, Qiao M, Kao A, Kölsch Y, Kumar A, Kunze VP, Rasys AM, Richardson R, Baier H, et al. 2023. Evolution of neuronal cell classes and types in the vertebrate retina. :2023.04.07.536039. Available from: https://www.biorxiv.org/content/10.1101/2023.04.07.536039v1

Hardie RC, Juusola M. 2015. Phototransduction in Drosophila. *Curr. Opin. Neurobiol.* [Internet] 34:37–45. Available from: https://www.sciencedirect.com/science/article/pii/S0959438815000173

Hattar S, Liao HW, Takao M, Berson DM, Yau KW. 2002. Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science* 295:1065–1070.

Heyland A, Croll R, Goodall S, Kranyak J, Wyeth R. 2014. Trichoplax adhaerens, an Enigmatic Basal Metazoan with Potential. In: Carroll DJ, Stricker SA, editors. Developmental Biology of the Sea Urchin and Other Marine Invertebrates: Methods and Protocols. Methods in Molecular Biology. Totowa, NJ: Humana Press. p. 45–61. Available from: https://doi.org/10.1007/978-1-62703-974-1_4

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* [Internet] 35:518–522. Available from: https://doi.org/10.1093/molbev/msx281

Horridge GA. 1964. Presumed photoreceptive cilia in a ctenophore. *Q. J. Microsc. Sci.* [Internet]. Available from: https://openresearch-repository.anu.edu.au/handle/1885/167542

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* [Internet] 33:1635–1638. Available from: https://doi.org/10.1093/molbev/msw046

Jékely G, Paps J, Nielsen C. 2015. The phylogenetic position of ctenophores and the origin(s) of nervous systems. *EvoDevo* [Internet] 6:1. Available from: https://doi.org/10.1186/2041-9139-6-1

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* [Internet] 30:1236–1240. Available from: https://doi.org/10.1093/bioinformatics/btu031

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* [Internet] 14:587–589. Available from: https://www.nature.com/articles/nmeth.4285

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* [Internet] 28:1947–1951. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3715

Kanehisa M, Sato Y, Kawashima M. 2021. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci.* [Internet] n/a. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4172

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* [Internet] 30:3059–3066. Available from: https://doi.org/10.1093/nar/gkf436

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* [Internet] 30:772–780. Available from: https://doi.org/10.1093/molbev/mst010

Koyanagi M, Ono K, Suga H, Iwabe N, Miyata T. 1998. Phospholipase C cDNAs from sponge and hydra: antiquity of genes involved in the inositol phospholipid signaling pathway1The nucleotide sequence data reported in this paper will appear in the DDBJ, EMBL and GenBank nucleotide sequence databases.1. *FEBS Lett.* [Internet] 439:66–70. Available from: https://www.sciencedirect.com/science/article/pii/S0014579398013398

Kozmik Z, Ruzickova J, Jonasova K, Matsumoto Y, Vopalensky P, Kozmikova I, Strnad H, Kawamura S, Piatigorsky J, Paces V, et al. 2008. Assembly of the cnidarian camera-type eye from vertebrate-like components. *Proc. Natl. Acad. Sci.* [Internet] 105:8989–8993. Available from: https://www.pnas.org/content/105/26/8989

Krishnan A, Mustafa A, Almén MS, Fredriksson R, Williams MJ, Schiöth HB. 2015. Evolutionary hierarchy of vertebrate-like heterotrimeric G protein families. *Mol. Phylogenet. Evol.* [Internet] 91:27–40. Available from: https://www.sciencedirect.com/science/article/pii/S1055790315001463

Kück P, Meusemann K. 2010. FASconCAT, Version 1.0, Zool. Forschungsmuseum A. Koenig, Germany, 2010.

Lagman D, Franzén IE, Eggert J, Larhammar D, Abalo XM. 2016. Evolution and expression of the phosphodiesterase 6 genes unveils vertebrate novelty to control photosensitivity. *BMC Evol. Biol.* [Internet] 16:124. Available from: https://doi.org/10.1186/s12862-016-0695-z

Lagman D, Sundström G, Ocampo Daza D, Abalo XM, Larhammar D. 2012. Expansion of transducin subunit gene families in early vertebrate tetraploidizations. *Genomics* [Internet] 100:203–211. Available from: https://www.sciencedirect.com/science/article/pii/S0888754312001358

Lamb TD. 2020. Evolution of the genes mediating phototransduction in rod and cone photoreceptors. *Prog. Retin. Eye Res.* [Internet] 76:100823. Available from: https://www.sciencedirect.com/science/article/pii/S1350946219301107

Lamb TD, Patel HR, Chuah A, Hunt DM. 2018. Evolution of the shut-off steps of vertebrate phototransduction. *Open Biol.* [Internet] 8:170232. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rsob.170232

Lee S-J, Xu H, Montell C. 2004. Rhodopsin kinase activity modulates the amplitude of the visual response in Drosophila. *Proc. Natl. Acad. Sci.* [Internet] 101:11874–11879. Available from: https://doi.org/10.1073/pnas.0402205101

Levy S, Elek A, Grau-Bové X, Menéndez-Bravo S, Iglesias M, Tanay A, Mass T, Sebé-Pedrós A. 2021. A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity. *Cell* [Internet] 184:2973-2987.e18. Available from: https://www.sciencedirect.com/science/article/pii/S0092867421004402

Leys SP, Degnan BM. 2001. Cytological Basis of Photoresponsive Behavior in a Sponge Larva. *Biol. Bull.* [Internet]. Available from: https://www.journals.uchicago.edu/doi/10.2307/1543611

Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* [Internet] 17:282–283. Available from: https://doi.org/10.1093/bioinformatics/17.3.282

Lukowski SW, Lo CY, Sharov AA, Nguyen Q, Fang L, Hung SS, Zhu L, Zhang T, Grünert U, Nguyen T, et al. 2019. A single-cell transcriptome atlas of the adult human retina. *EMBO J.* [Internet] 38:e100811. Available from: https://doi.org/10.15252/embj.2018100811

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* [Internet] 161:1202–1214. Available from: https://www.cell.com/cell/abstract/S0092-8674(15)00549-8

Maddison W, Maddison D. 2008. Mesquite: A modular system for evolutionary analysis. *Evolution* 62:1103–1118.

Maldonado M, Durfort M, McCarthy DA, Young CM. 2003. The cellular basis of photobehavior in the tufted parenchymella larva of demosponges. *Mar. Biol.*

[Internet] 143:427–441. Available from: https://doi.org/10.1007/s00227-003-1100-1

Mikami K. 2014. Structural divergence and loss of phosphoinositide-specific phospholipase C signaling components during the evolution of the green plant lineage: implications from structural characteristics of algal components. *Front. Plant Sci.* [Internet] 5:380. Available from: https://www.frontiersin.org/article/10.3389/fpls.2014.00380

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* [Internet] 37:1530–1534. Available from: https://doi.org/10.1093/molbev/msaa015

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Mol. Biol. Evol.* [Internet] 37:2763–2774. Available from: https://doi.org/10.1093/molbev/msaa141

Müller WEG, Schröder HC, Markl JS, Grebenjuk VA, Korzhev M, Steffen R, Wang X. 2013. Cryptochrome in Sponges: A Key Molecule Linking Photoreception with Phototransduction. *J. Histochem. Cytochem.* [Internet] 61:814–832. Available from: https://doi.org/10.1369/0022155413502652

Mushegian A, Gurevich VV, Gurevich EV. 2012. The Origin and Evolution of G Protein-Coupled Receptor Kinases. *PLOS ONE* [Internet] 7:e33806. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033806

Nilsson D-E. 2009. The evolution of eyes and visually guided behaviour. *Philos. Trans. R. Soc. B Biol. Sci.* [Internet] 364:2833–2847. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2009.0083

Nilsson D-E. 2013. Eye evolution and its functional basis. *Vis. Neurosci.* [Internet] 30:5–20. Available from: https://www.cambridge.org/core/journals/visual-neuroscience/article/eye-evolution-and-its-functional-basis/E632F655150C8D0E7367566CC99F4717

Nordström K, Wallén null, Seymour J, Nilsson D. 2003. A simple visual system without neurons in jellyfish larvae. *Proc. R. Soc. Lond. B Biol. Sci.* [Internet] 270:2349–2354. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2003.2504

Orban T, Palczewski K. 2016. Structure and Function of G-Protein-Coupled Receptor Kinases 1 and 7. In: Gurevich VV, Gurevich EV, Tesmer JJG, editors. G Protein-Coupled Receptor Kinases. New York, NY: Springer New York. p. 25–43. Available from: https://doi.org/10.1007/978-1-4939-3798-1_2

Özel MN, Simon F, Jafari S, Holguera I, Chen Y-C, Benhra N, El-Danaf RN, Kapuralin K, Malin JA, Konstantinides N, et al. 2021. Neuronal diversity and convergence in a visual system developmental atlas. *Nature* [Internet] 589:88–95. Available from: https://www.nature.com/articles/s41586-020-2879-3

Paganos P, Voronov D, Musser JM, Arendt D, Arnone MI. 2021. Single-cell RNA sequencing of the Strongylocentrotus purpuratus larva reveals the blueprint of major cell types and nervous system of a non-chordate deuterostome.Tessmar-

Raible K, Bronner ME, Martinez Serra P, Revilla-i-Domingo R, Hinman V, editors. *eLife* [Internet] 10:e70416. Available from: https://doi.org/10.7554/eLife.70416

Palczewski K, Kiser PD. 2020. Shedding new light on the generation of the visual chromophore. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 117:19629–19638. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7443880/

Passamaneck YJ, Furchheim N, Hejnol A, Martindale MQ, Lüter C. 2011. Ciliary photoreceptors in the cerebral eyes of a protostome larva. *EvoDevo* [Internet] 2:6. Available from: https://doi.org/10.1186/2041-9139-2-6

Piatigorsky J, Kozmik Z. 2004. Cubozoan jellyfish: an Evo/Devo model for eyes and other sensory systems. *Int. J. Dev. Biol.* [Internet] 48:719–729. Available from: http://www.ijdb.ehu.es/web/paper/041851jp

Picciani N, Kerlin JR, Sierra N, Swafford AJM, Ramirez MD, Roberts NG, Cannon JT, Daly M, Oakley TH. 2018. Prolific Origination of Eyes in Cnidaria with Co-option of Non-visual Opsins. *Curr. Biol.* [Internet] 28:2413-2419.e4. Available from: https://www.sciencedirect.com/science/article/pii/S0960982218306912

Plachetzki DC, Fong CR, Oakley TH. 2010. The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway. *Proc. R. Soc. B Biol. Sci.* [Internet] 277:1963–1969. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2009.1797

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* [Internet] 33:3454–3460. Available from: https://doi.org/10.1093/bioinformatics/btx439

Provencio I, Rodriguez IR, Jiang G, Hayes WP, Moreira EF, Rollag MD. 2000. A Novel Human Opsin in the Inner Retina. *J. Neurosci.* [Internet] 20:600–605. Available from: https://www.jneurosci.org/content/20/2/600

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res.* [Internet] 33:W116–W120. Available from: https://doi.org/10.1093/nar/gki442

Rebecchi MJ, Pentyala SN. 2000. Structure, Function, and Control of Phosphoinositide-Specific Phospholipase C. *Physiol. Rev.* [Internet] 80:1291–1335. Available from: https://journals.physiology.org/doi/full/10.1152/physrev.2000.80.4.1291

Rivera AS, Ozturk N, Fahey B, Plachetzki DC, Degnan BM, Sancar A, Oakley TH. 2012. Blue-light-receptive cryptochrome is expressed in a sponge eye lacking neurons and opsin. *J. Exp. Biol.* [Internet] 215:1278–1286. Available from: https://doi.org/10.1242/jeb.067140

Rollag MD, Berson DM, Provencio I. 2003. Melanopsin, Ganglion-Cell Photoreceptors, and Mammalian Photoentrainment. *J. Biol. Rhythms* [Internet] 18:227–234. Available from: https://doi.org/10.1177/0748730403018003005

Ryan K, Lu Z, Meinertzhagen IA. 2016. The CNS connectome of a tadpole larva of Ciona intestinalis (L.) highlights sidedness in the brain of a chordate

sibling.Marder E, editor. *eLife* [Internet] 5:e16962. Available from: https://doi.org/10.7554/eLife.16962

Schnitzler CE, Pang K, Powers ML, Reitzel AM, Ryan JF, Simmons D, Tada T, Park M, Gupta J, Brooks SY, et al. 2012. Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. *BMC Biol.* [Internet] 10:107. Available from: https://doi.org/10.1186/1741-7007-10-107

Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient gene linkages support ctenophores as sister to other animals. *Nature* [Internet]:1–8. Available from: https://www.nature.com/articles/s41586-023-05936-6

Schwarz G. 1978. Estimating the Dimension of a Model. *Ann. Stat.* [Internet] 6:461–464. Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-2/Estimating-the-Dimension-of-a-Model/10.1214/aos/1176344136.full

Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, Amit I, Hejnol A, Degnan BM, Tanay A. 2018. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* 2:1176–1188.

Sebé-Pedrós A, Saudemont B, Chomsky E, Plessier F, Mailhé M-P, Renno J, Loe-Mie Y, Lifshitz A, Mukamel Z, Schmutz S, et al. 2018. Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* [Internet] 173:1520-1534.e20. Available from: https://www.sciencedirect.com/science/article/pii/S0092867418305968

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* [Internet] 13:2498–2504. Available from: https://genome.cshlp.org/content/13/11/2498

Sharma S, Wang W, Stolfi A. 2019. Single-cell transcriptome profiling of the Ciona larval brain. *Dev. Biol.* [Internet] 448:226–236. Available from: https://www.sciencedirect.com/science/article/pii/S0012160618300137

Shen W-K, Chen S-Y, Gan Z-Q, Zhang Y-Z, Yue T, Chen M-M, Xue Y, Hu H, Guo A-Y. 2023. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.* 51:D39–D45.

Shichida Y, Matsuyama T. 2009. Evolution of opsins and phototransduction. *Philos. Trans. R. Soc. B Biol. Sci.* [Internet] 364:2881–2895. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2781858/

Siebert S, Farrell JA, Cazet JF, Abeykoon Y, Primack AS, Schnitzler CE, Juliano CE. 2019. Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *Science* [Internet] 365:eaav9314. Available from: https://www.science.org/doi/10.1126/science.aav9314

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy

orthologs. *Bioinformatics* [Internet] 31:3210–3212. Available from: https://doi.org/10.1093/bioinformatics/btv351

Smith CL, Varoqueaux F, Kittelmann M, Azzam RN, Cooper B, Winters CA, Eitel M, Fasshauer D, Reese TS. 2014. Novel Cell Types, Neurosecretory Cells, and Body Plan of the Early-Diverging Metazoan Trichoplax adhaerens. *Curr. Biol.* [Internet] 24:1565–1572. Available from: https://www.sciencedirect.com/science/article/pii/S0960982214006113

Suh P-G, Park J-I, Manzoli L, Cocco L, Peak JC, Katan M, Fukami K, Kataoka T, Yun S, Ryu SH. 2008. Multiple roles of phosphoinositide-specific phospholipase C isozymes. *BMB Rep.* 41:415–434.

Tamm SL. 2016. Novel Structures Associated with Presumed Photoreceptors in the Aboral Sense Organ of Ctenophores. *Biol. Bull.* 231:97–102.

Terakita A. 2005. The opsins. *Genome Biol.* [Internet] 6:213. Available from: https://doi.org/10.1186/gb-2005-6-3-213

Tsutsui K, Minami J, Matsushita O, Katayama S, Taniguchi Y, Nakamura S, Nishioka M, Okabe A. 1995. Phylogenetic analysis of phospholipase C genes from Clostridium perfringens types A to E and Clostridium novyi. *J. Bacteriol.* [Internet] 177:7164–7170. Available from: https://journals.asm.org/doi/abs/10.1128/jb.177.24.7164-7170.1995

Ullrich-Lüter EM, Dupont S, Arboleda E, Hausen H, Arnone MI. 2011. Unique system of photoreceptors in sea urchin tube feet. *Proc. Natl. Acad. Sci. U. S. A.* 108:8367–8372.

Valencia JE, Feuda R, Mellott DO, Burke RD, Peter IS. 2021. Ciliary photoreceptors in sea urchin larvae indicate pan-deuterostome cell type conservation. *BMC Biol.* [Internet] 19:257. Available from: https://doi.org/10.1186/s12915-021-01194-y

Varoqueaux F, Williams EA, Grandemange S, Truscello L, Kamm K, Schierwater B, Jékely G, Fasshauer D. 2018. High Cell Diversity and Complex Peptidergic Signaling Underlie Placozoan Behavior. *Curr. Biol.* [Internet] 28:3495-3501.e2. Available from: https://www.sciencedirect.com/science/article/pii/S0960982218311977

Vöcking O, Macias-Muñoz A, Jaeger S, Oakley TH. 2022. Deep Diversity: Extensive Variation in the Components of Complex Visual Systems across Animals. Available from: https://www.preprints.org/manuscript/202209.0432/v1

Vopalensky P, Kozmik Z. 2009. Eye evolution: common use and independent recruitment of genetic components. *Philos. Trans. R. Soc. B Biol. Sci.* [Internet] 364:2819–2832. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2009.0079

Wang T, Montell C. 2007. Phototransduction and retinal degeneration in Drosophila. *Pflüg. Arch. - Eur. J. Physiol.* [Internet] 454:821–847. Available from: https://doi.org/10.1007/s00424-007-0251-1

Wang X, Liu Y, Li Z, Gao X, Dong J, Yang M. 2020. Expression and evolution of the phospholipase C gene family in Brachypodium distachyon. *Genes Genomics*

[Internet] 42:1041–1053. Available from: https://doi.org/10.1007/s13258-020-00973-1

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* 35:543–548.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017. Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.* [Internet] 1:1737–1746. Available from: https://www.nature.com/articles/s41559-017-0331-3

Widjaja-Adhi MAK, Golczak M. 2020. The molecular aspects of absorption and metabolism of carotenoids and retinoids in vertebrates. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1865:158571.

Wong E, Anggono V, Williams SR, Degnan SM, Degnan BM. 2022. Phototransduction in a marine sponge provides insights into the origin of animal vision. *iScience* [Internet] 25:104436. Available from: https://www.sciencedirect.com/science/article/pii/S2589004222007076

Yau K-W, Hardie RC. 2009. Phototransduction Motifs and Variations. *Cell* [Internet] 139:246–264. Available from: https://www.sciencedirect.com/science/article/pii/S0092867409012446

Zuber ME, Gestri G, Viczian AS, Barsacchi G, Harris WA. 2003. Specification of the vertebrate eye by a network of eye field transcription factors. *Development* [Internet] 130:5155–5167. Available from: https://doi.org/10.1242/dev.00723

Chapter 4

The Evolution of Retinol Metabolism and Implications for the Origin of Vision

Abstract

Vision in animals fundamentally relies on a light-sensitive molecule, an opsin protein bound to a derivative of vitamin A, typically 11-cis-retinal. Upon light absorption, 11-cis-retinal undergoes a conformational change to its trans-state. To regain light sensitivity, the 11-cis configuration must be restored, a process dependent on retinol metabolism, which produces 11-cis-retinal from dietary vitamin A. Since opsins are unable to respond to light in absence of the 11-cis-retinal, understanding the evolution of retinol metabolism offers insights into the origin of vision itself. Yet, the evolution and diversity of enzymes integral to this pathway remain elusive.

The aim of this chapter was to investigate the evolution of the retinol metabolism pathway by identifying the orthogroups—phylogenetically defined gene families—its enzymes belong to and characterizing their evolutionary history across diverse eukaryotic lineages. The results identified 12 overarching orthogroups encompassing the enzymes involved in the retinol metabolism. These orthogroups are generally very ancient and span the eukaryotic domain. Phylogenetic analyses uncovered intricate substructures within each orthogroup, revealing multiple sub-families. Intriguingly, the orthogroups containing some of most quintessential enzymes of the retinol metabolism (e.g., BCMO1 and RPE65) exhibit a pattern where the specific sub-family engaged in the pathway is found exclusively in animals, despite the wider eukaryotic distribution of the overarching orthogroup. Such findings allude to animal-specific expansions of these gene families, concurrent with the emergence of vision.

Introduction

The retinol metabolism comprises a series of enzymatic reactions that convert dietary vitamin A into various bioactive compounds, primarily retinal for vision and retinoic acid for gene regulation, ensuring the proper functioning of visual processes and other physiological roles in the body (Blomhoff and Blomhoff 2006; Dewett, Lam-Kamath, et al. 2021). Retinol (Vitamin A_1) is an essential micronutrient derived primarily from diet. It can be obtained directly from animal sources as retinyl esters or indirectly from plant sources as pro-vitamin A carotenoids, which are then converted into retinol in the body (Trifiletti 2014). In turn, retinol can be esterified to retinyl ester by the enzyme lecithin retinol acyltransferase (LRAT) allowing for its storage (Batten et al. 2004). When needed, retinyl ester is hydrolysed back to retinol (Moiseyev et al. 2005). Retinol is oxidized to retinal by retinol dehydrogenases (RDHs). Several other enzymes are involved in various steps of the retinol metabolism pathway as summarized in Figure 4.1, which is based on what is known about the pathway according to the KEGG Pathway Database (Kanehisa et al. 2021). In addition, Table 4.1 provides a comprehensive list of all enzymes. As several of them are involved in other biological process, in Table 4.1, they are ranked by the number of pathways they participate in according to KEGG. Involvement in one or few pathways serves as an indicator of enzyme specificity to the retinol metabolism, as opposed to multifunctional enzymes involved in numerous pathways.

Retinal, particularly 11-cis-retinal, plays a crucial role in vision (Palczewski and Kiser 2020). 11-cis-retinal binds to the protein opsin in photoreceptor cells forming rhodopsin. Upon absorbing a photon, 11-cis-retinal is isomerized to all-trans-retinal, leading to a conformational change in opsin, and initiating a cascade of events called phototransduction (Hardie and Juusola 2015; Lamb 2020) (see Chapter 3). After light exposure, all-trans-retinal is reduced to all-trans-retinol and then converted back to 11-cis-retinal through a series of enzymatic reactions. This part of the visual cycle is essential as it ensures the retina's responsiveness to light (Palczewski and Kiser 2020). The regulation of the metabolic steps ensures sufficient 11-cis-retinal availability and prevents toxic build-up of intermediates. Additionally, retinal can be further oxidized to retinoic acid by retinaldehyde dehydrogenase (RALDH1). Retinoic acid serves as a signalling

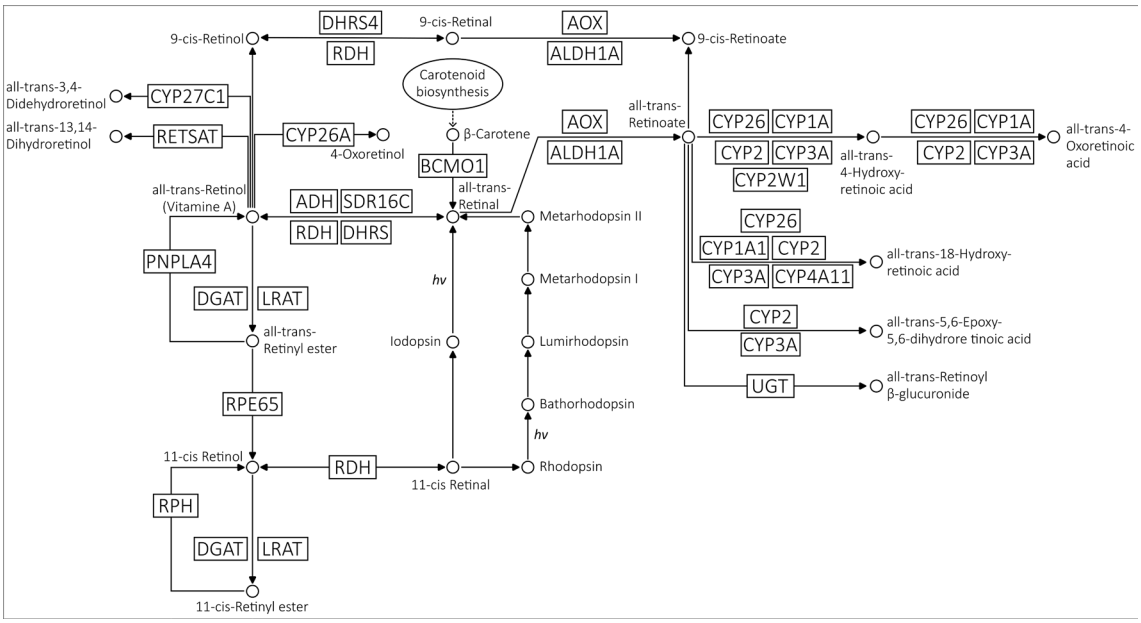


Figure 4.2. Retinol metabolism pathway. The pathway used as reference in this study is based on the KEGG map00830 (Kanehisa et al. 2021).

molecule that regulates gene expression and is critical for numerous developmental processes (Blomhoff and Blomhoff 2006). The retinol metabolism, particularly as it relates to vision, has been primarily studied in vertebrates, especially mammals, with mouse (*Mus musculus*) and human being the most extensively characterized due to their relevance in medical research (Trifiletti 2014; Widjaja-Adhi and Golczak 2020). Some aspects of retinol metabolism have been studied in the invertebrate model *Drosophila melanogaster*, where vitamin A deficiency hampers the formation of the retinal and overall disrupts the visual system (Dewett, Labaf, et al. 2021; Dewett, Lam-Kamath, et al. 2021). Outside of animals, carotenoid biosynthesis pathways, producing retinol precursors such as beta-carotene, have received more attention than the retinol metabolism itself, especially in plants (Hirschberg 2001).

Given the importance of retinol metabolism, it is compelling to explore its evolutionary history and potential diversity outside of traditional model organisms, especially in the wider context of the evolution of vision. Hence, the work presented in this chapter aimed to unravel this intricate history. The initial step was to identify the genetic components involved and determine their evolutionary relationships to answer questions such as: Do the gene families belong to overarching orthogroups? How closely related are they? The subsequent objective was to uncover the distribution of these components across the animal kingdom and, more broadly, within eukaryotes, to pinpoint the specific point in

time when all the components came into place. The final endeavour was to delineate the main evolutionary events characterizing each orthogroup, to discern, for instance, if certain gene families have undergone a greater number of evolutionary events and contextualizing them within the evolutionary tree of life.

Table 4.1. Enzymes involved in retinol metabolism listed in order of number of pathways they are involved in according to KEGG, as a measure of their specificity to the retinol metabolism pathway.

Gene Family	Kegg map00830			No Kegg Pathways
	Entry	Symbol	Name	
RETSAT	K09516	RETSAT	all-trans-retinol 13,14-reductase	1
RPH	EC3.1.1.63	?	11-cis-retinyl-palmitate hydrolase	1
PNPLA4	K11157	PNPLA4	patatin-like phospholipase domain-containing protein 4	2
RDH	K00061	RDH5	11-cis-retinol dehydrogenase	2
	K11150	RDH8	retinol dehydrogenase 8	2
	K11151	RDH10	retinol dehydrogenase 10	2
	K11152	RDH11	retinol dehydrogenase 11	3
	K11153	RDH12	retinol dehydrogenase 12	3
	K11161	RDH13	retinol dehydrogenase 13	3
ALDH1	K07249	ALDH1A	retinal dehydrogenase	2
RPE65	K11158	RPE65	retinoid isomerohydrolase / lutein isomerase	2
BCMO1	K00515	BCMO1	beta-carotene 15,15'-dioxygenase	3
LRAT	K00678	LRAT	phosphatidylcholine-retinol O-acyltransferase	3
DHRS	K11148	DHRS4L2	dehydrogenase/reductase SDR family member 4-like protein 2	2
	K11149	DHRS9	dehydrogenase/reductase SDR family member 9	2
	K11146	DHRS3	short-chain dehydrogenase/reductase 3	3
	K11147	DHRS4	dehydrogenase/reductase SDR family member 4	3
	K15734	SDR16C5	all-trans-retinol dehydrogenase (NAD+)	3
	K13369	HSD17B6	17beta-estradiol 17-dehydrogenase / all-trans-retinol dehydrogenase (NAD+)	4
DGAT	K11156	DGAT2L4	diacylglycerol O-acyltransferase 2-like protein 4	3
	K11155	DGAT1	diacylglycerol O-acyltransferase 1	4
CYP	K17951	CYP27C1	all-trans-retinol 3,4-desaturase	1
	K07411	CYP2A	cytochrome P450 family 2 sub-family A	2
	K07423	CYP2W1	cytochrome P450 family 2 sub-family W1	2
	K07437	CYP26A	cytochrome P450 family 26 sub-family A	2
	K12664	CYP26B	cytochrome P450 family 26 sub-family B	2
	K12665	CYP26C	cytochrome P450 family 26 sub-family C	2
	K07420	CYP2S1	cytochrome P450 family 2 sub-family S1	3
	K17691	CYP3A7	cytochrome P450 family 3 sub-family A7	4
	K17720	CYP2C18	cytochrome P450 family 2 sub-family C18	4
	K07412	CYP2B	cytochrome P450 family 2 sub-family B	5
	K07424	CYP3A	cytochrome P450 family 3 sub-family A	6
	K17690	CYP3A5	cytochrome P450 family 3 sub-family A5	6
	K07425	CYP4A	long-chain fatty acid omega-monooxygenase	7
	K17709	CYP2B6	cytochrome P450 family 2 sub-family B6	7
	K07413	CYP2C	cytochrome P450 family 2 sub-family C	8
	K17683	CYP2A6	cytochrome P450 family 2 sub-family A6	8
	K17718	CYP2C8	cytochrome P450 family 2 sub-family C8	8
	K17719	CYP2C9	cytochrome P450 family 2 sub-family C9	9
	K07408	CYP1A1	cytochrome P450 family 1 sub-family A1	10
	K17689	CYP3A4	cytochrome P450 family 3 sub-family A4	10
	K07409	CYP1A2	cytochrome P450 family 1 sub-family A2	12
AOX	K00157	AOX	aldehyde oxidase	10
ADH	K13980	ADH4	alcohol dehydrogenase 4	9
	K13951	ADH1_7	alcohol dehydrogenase 1/7	10
	K13952	ADH6	alcohol dehydrogenase 6	10
	K00001	adh	alcohol dehydrogenase	13
	K13953	adhP	alcohol dehydrogenase, propanol-preferring	13
	K00121	ADH5	alcohol dehydrogenase 5	16
UGT	K00699	UGT	glucuronosyltransferase	14

Results and Discussion

Enzymes involved in retinol metabolism belong to 12 major orthogroups.

To gain insights into the evolution of retinol metabolism, it is essential to trace the evolutionary history of each of the enzymes in the pathway. For this I used as reference the pathway described by KEGG (Kanehisa et al. 2021) (Figure 4.1 and Table 4.1) and explored the genes encoding these enzymes across 101 species spanning all major eukaryotic groups (Table 4.2 and Extended Table 4.2). KEGG ortholog lists (Kanehisa 2019) for each enzyme were used as starting point for the data mining (see more details in the Methods section of this Chapter). It is worth noting that the only enzyme from the KEGG pathway excluded from this analysis was RPH (11-cis-retinyl-palmitate hydrolase) (Figure 4.1 and Table 4.1). Despite its hypothesized role in hydrolysing stored 11-cis-retinyl esters to 11-cis retinol is pertinent to vision (Blaner et al. 1984; Blaner et al. 1987), there is a significant knowledge gap surrounding this putative enzyme. The human gene encoding it remains unidentified, and KEGG does not list any orthologs for it. Given the nebulous nature of this enzyme, this study chose to prioritize better-understood enzymes, including RPE65 that catalyses the extremely similar reaction of hydrolysing all-trans-retinyl esters to 11-cis retinol (Moiseyev et al. 2005).

Although many enzymes partake in the pathway, some might belong to a same larger gene family. Therefore, to study their evolution, the initial task was to identify their respective orthogroup – a collection of orthologs and paralogs that originated from the same initial gene duplication. An orthogroup can be considered as a phylogenetically defined gene family. Orthogroup inference methods often rely on computing sequence similarity scores amongst sequences as a measure of protein distances and then using these scores for clustering the sequences (e.g., OrthoMCL (Li et al. 2003)). Here, two alternative software for orthogroup inference were used to independently infer orthogroups (see details in Methods). The first was OrthoFinder that implements a method that eliminates gene length bias during similarity score assessment (Emms and Kelly 2015) and uses a

Table 4.2. List of species used in this study with respective proteome BUSCO scores.

Clades							Species	% Complete BUSCOs (tot) (eukaryota_odb10)
Amorphea	Obazoa	Opisthokonta	Holozoa	Metazoa	Bilateria	Mollusca	*Octopus bimaculoides*	92.90%
							Lottia gigantea	96.50%
						Brachiopoda	*Lingula unguis*	98.40%
						Annelida	*Capitella teleta*	94.20%
							Helobdella robusta	97.30%
						Bryozoa	*Bugula neritina*	54.50%
						Acoelomorpha	*Hofstenia miamia*	83.60%
						Platyhelminthes	*Echinococcus multilocularis*	90.20%
						Nematoda	*Caenorhabditis elegans*	97.70%
							Loa loa	96.10%
						Rotifera	*Brachionus plicatilis*	90.60%
						Tardigrada	*Ramazzotius varieornatus*	92.50%
						Arthropoda	*Daphnia pulex*	97.70%
							Drosophila melanogaster	100.00%
							Calanus glacialis	96.80%
						Vertebrate	*Homo sapiens*	100.00%
							Mus musculus	100.00%
							Danio rerio	99.60%
							Eptatretus burgeri	85.90%
						Urochordata	*Ciona intestinalis*	97.30%
							Oikopleura dioica	80.70%
						Cephalochordata	*Branchiostoma belcheri*	96.90%
						Echinodermata	*Acanthaster planci*	91.40%
							Strongylocentrotus purpuratus	96.00%
						Hemichordata	*Saccoglossus kowalevskii*	94.10%
					Cnidaria		*Fungia scutaria*	84.40%
							Montastraea cavernosa	80.40%
							Madracis auretenra	68.30%
							Stylophora pistillata	85.10%
							Astreopora sp	69.80%
							Porites australiensis	81.60%
							Acropora digitifera	42.80%
							Anthopleura elegantissima	62.00%
							Nematostella vectensis	93.40%
							Gorgonia ventalina	56.50%
							Clytia hemisphaerica	84.70%
							Hydra magnapapillata	70.90%
							Aurelia sp	67.40%
							Thelohanellus kitauei	29.40%
					Placozoa		*Trichoplax adhaerens (spH2)*	96.10%
							Hoilungia hongkongensis	96.80%
					Ctenophora		*Mnemiopsis leidyi*	83.60%
							Pleurobrachia bachei	47.50%
					Porifera		*Leucosolenia complicata*	94.90%
							Sycon ciliatum	97.30%
							Stylissa carteri	43.60%
							Oscarella pearsei	94.90%
							Amphimedon queenslandica	92.50%
							Haliclona tubifera	75.30%
							Ephydatia muelleri	93.40%
				Choanoflagellata			*Acanthoeca spectabilis*	93.30%
							Helgoeca nana	94.10%
							Diaphanoeca grandis	92.20%
							Didymoeca costata	94.90%
							Choanoeca perplexa	92.10%
							Monosiga brevicollis	78.80%
							Mylnosiga fluctuans	93.30%
							Salpingoeca kvevrii	94.90%
				Ichthyosporea			*Amoebidium parasiticum*	89.90%
							Sphaeroforma arctica	63.20%
							Sphaerothecum destruens	68.60%
				Pluriformea			*Corallochytrium limacisporum*	90.20%
				Filastrea			*Capsaspora owczarzaki*	93.70%
			Nucletmycea (Holomycota)	Fungi			*Fusarium oxysporum*	97.30%
							Saccharomyces cerevisiae	93.30%
							Ustilago maydis	98.80%
							Mortierella elongata	98.40%
							Rhizopus microsporus	97.30%
							Spizellomyces punctatus	94.90%
				Rotosphaerida			*Parvularia atlantis*	75.70%
							Fonticula alba	71.00%
		Apusomonadida					*Thecamonas trahens*	81.60%
		Breviata					*Pygsuia biforma*	62.40%
	Amoebozoa						*Dictyostelium discoideum*	94.10%
							Vermamoeba vermiformis	89.00%
							Cunea spBSH02190019	76.50%
	CRuMs	Collodictyonidae					*Diphylleia rotans*	90.60%
		Rigifilida					*Rigifila ramosa*	86.30%
Incertae sedis	Malawimonadidae						*Gefionella okellyi*	78.50%
Diaphoretickes	Archaeplastida			Chloroplastida			*Arabidopsis thaliana*	99.60%
							Marchantia polymorpha	99.60%
							Chlamydomonas reinhardtii	91.80%
							Chloropicon primus	89.80%
				Glaucophyta			*Gloeochaete wittrockiana*	85.50%
				Rhodophyta			*Porphyridium purpureum*	72.60%
							Galdieria sulphuraria	77.30%
				Rhodelphis			*Rhodelphis marinus*	89.80%
	Haptista			Haptophyta			*Prymnesium parvum*	71.40%
							Phaeocystis globosa	75.30%
				Centroplasthelida			*Choanocystis spHF7*	88.20%
	Cryptista			Cryptophyceae			*Baffinella spCCMP2293*	78.90%
							Guillardia theta	79.60%
	Sar			Stramenopiles			*Aphanomyces astaci*	95.60%
							Phytophthora infestans	90.20%
				Rhizaria			*Plasmodiophora brassicae*	89.00%
				Alveolata			*Colponemidia spColp10*	93.30%
							Tetrahymena thermophila	76.40%
	Telonemia						*Telonemia subtile*	82%
	Excavata			Euglenozoa			*Euglena longa*	83.50%
				Heterolobosea			*Pharyngomonas kirbyi*	85.10%
				Jakobida			*Andalucia godoyi*	79.20%

>90%
>80%
>70%
>60%
>50%
<50%

phylogenetic framework to detect orthologs (Emms and Kelly 2019); the second was Broccoli that uses phylogenetic relationships instead of protein distances for clustering sequences and then applies machine learning algorithms to extract orthologous relationships from this network (Derelle et al. 2020). By comparing results from these distinct strategies, the chances of comprehensively identifying orthogroups for retinol metabolism enzymes was enhanced.

OrthoFinder identified a total of 50 orthogroups, while Broccoli provided 58. After annotating the orthogroups and filtering out the ones not involved in the retinol metabolism (see Methods), we were left with 14 OrthoFinder and 21 Broccoli orthogroups (Figure 4.2). Results were compared by assessing the percentage of shared sequences between OrthoFinder and Broccoli orthogroups (see Methods for details). Generally, there is substantial agreement between OrthoFinder and Broccoli results, with many orthogroups displaying one-to-one correspondence. However, while OrthoFinder yielded fewer, larger orthogroups, Broccoli in some cases produced more and smaller ones. As a result, some gene families were fragmented into multiple smaller orthogroups exclusively in Broccoli's output. Collectively, the OrthoFinder and Broccoli results delineated 12 orthogroups encompassing retinol metabolism enzymes (Table 4.3). While the primary purpose of the orthogroup inference step was to identify gene families to investigate further with phylogenetic analyses, it also provided some preliminary insights into the evolution of some of the enzymes involved in retinol metabolism. For example, both OrthoFinder and Broccoli place DGAT1 and DGAT2L4 into distinct orthogroups. Additionally, RDH and DHRS enzymes, sub-families of a larger group, display a complex substructure, suggesting intricate phylogenetic relationships.

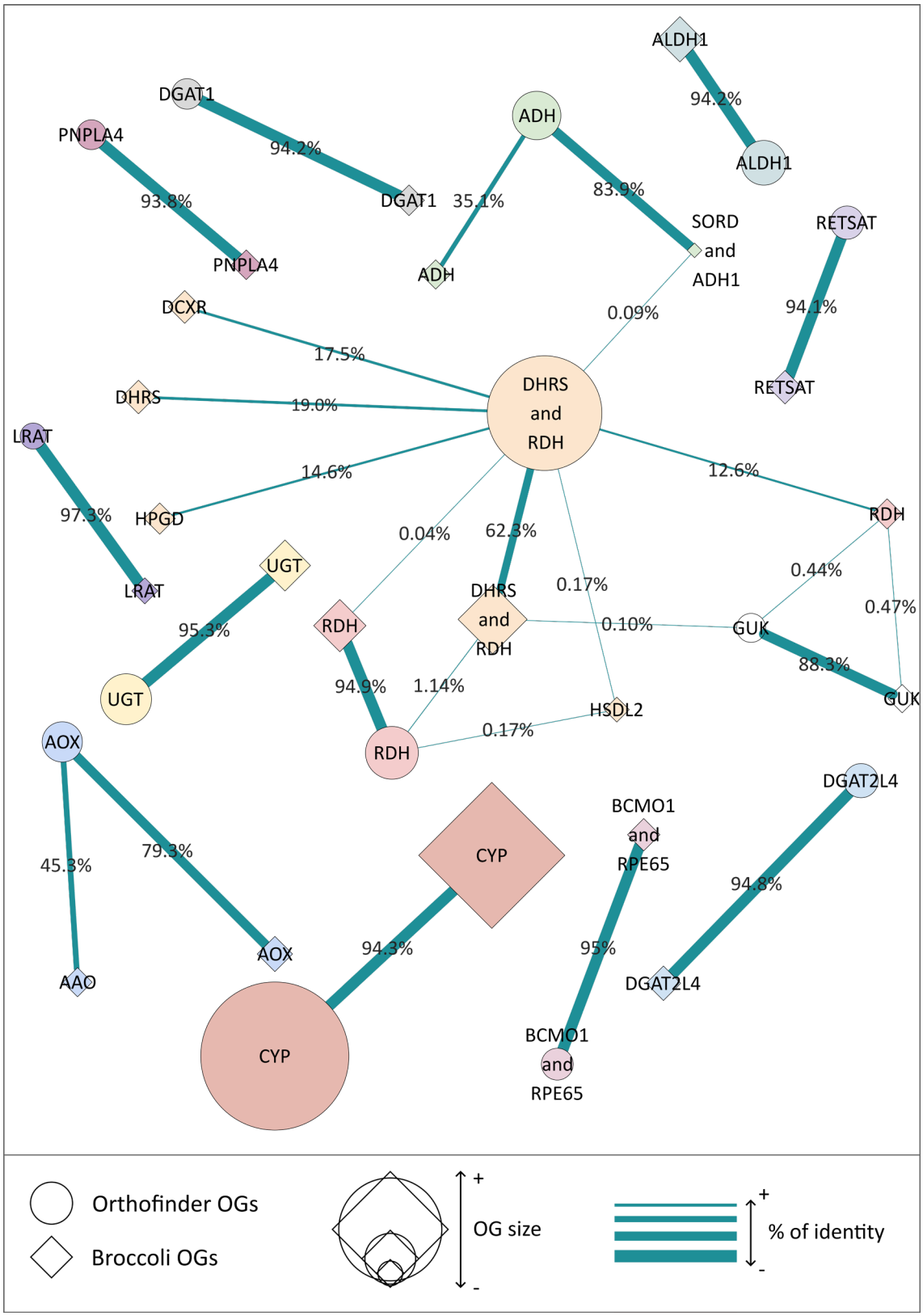


Figure 4.2. Orthogroup inference analysis. Orthogroups inferred from two different software (OrthoFinder and Broccoli) are compared.

Table 4.3. Summary of the comparison between OrthoFinder and Broccoli orthogroup inference results. In the last column, the final orthogroups used for phylogenetic analyses are shown.

Kegg Groups	Orthofinder Orthogroups (OOG)	Broccoli Orthogroups (BOG)	Consensus Groups
RETSAT	OOG16-RETSAT	BOG37-RETSAT	**RETSAT**
PNPLA4	OOG20-PNPLA4	BOG51-PNPLA4	**PNPLA4**
ALDH1	OOG6-ALDH1	BOG33-ALDH1	**ALDH1**
BCMO1 RPE65	OOG17-BCMO1-RPE65	BOG49-BCMO1-RPE65	**BCMO1-RPE65**
LRAT	OOG27-LRAT	BOG52-LRAT	**LRAT**
DHRS	OOG1-DHRS-RDH	BOG18-DHRS BOG3-DHRS-RDH BOG17-HPGD BOG22-DCXR	**RDH-DHRS**
		BOG19 -HSDL2	
RDH	OOG3-RDH	BOG5-RDH BOG15-RDH	
DGAT	OOG19-DGAT1	BOG27-DGAT1	**DGAT1**
	OOG14-DGAT2L4	BOG46-AWAT2	**DGAT2L4**
CYP	OOG0-CYP	BOG-CYP	**CYP**
AOX	OOG7-AOX	BOG38-AOX BOG39-AAO3	**AOX**
ADH	OOG5-ADH	BOG14-ADH BOG11-SORD-ADH1	**ADH**
UGT	OOG4-UGT	BOG30-UGT	**UGT**

Reconstructing phylogenetic histories of retinol metabolism orthogroups.

All retinol metabolism related orthogroups were further examined with phylogenetic analyses to understand the details of their evolutionary histories. After constructing phylogenetic trees (see Methods), two distinct but complementary approaches were applied for analysing each orthogroup tree. The first approach employs the software Possvm (Grau-Bové and Sebé-Pedrós 2021), which identifies orthologs within the gene tree through species overlap algorithms, defines sub-orthogroups within the primary orthogroup, and annotates the tree based on these sub-orthogroups. The second employs GeneRax (Morel et al. 2020), which reconciles the gene tree to a species tree using a maximum-likelihood framework. Possvm offers the advantage of swiftly annotating large trees, facilitating their interpretation. As it infers orthologs using implicit taxonomic information from the gene tree, it eliminates the need for a species tree and avoids potential biases from contentious species relationships. GeneRax, in contrast, delivers a precise reconciled tree detailing speciation, duplication, and loss events at each node. However, it demands more computation time and necessitates a species tree. The detailed results for each orthogroup, presented in order of specificity to the retinol metabolism, are described below.

RETSAT

The Retinol Saturase (RETSAT) enzyme catalyses the reaction that saturates the 13-14 double bond of all-trans-retinol to produce all-trans-13,14-dihdriretinol (Moise et al. 2004) (Figure 4.1). This enzyme appears to be involved only in retinol metabolism according to the KEGG Database (Table 4.1), meaning it is very specific to this pathway.

The orthogroups identified for RETSAT by OrthoFinder and Broccoli present a clear one-to-one relationship with high degree of identity (Figure 4.2), indicating no mixture with any other orthogroup examined. The merged RETSAT orthogroup contained 338 sequences distributed throughout all major eukaryotic clades (Figure 4.3A).

Phylogenetic analysis identified a monophyletic clade containing RETSAT genes from various species of eukaryotes, as well as other clades of related enzymes (Figure 4.3 B and C). Ortholog sorting with Possvm identified 7 orthogroups within the RETSAT family, with one orthogroup containing RETSAT, PYRD2 (Pyridine Nucleotide-Disulphide Oxidoreductase Domain 2) and CRT enzymes that are involved in carotenoid

metabolism (Figure 4.3B). Gene tree to species tree reconciliation with GeneRax confirmed the overall topology and revealed a high number of evolutionary events (especially losses) in proportion to the size of the orthogroup (Figure 4.3C).

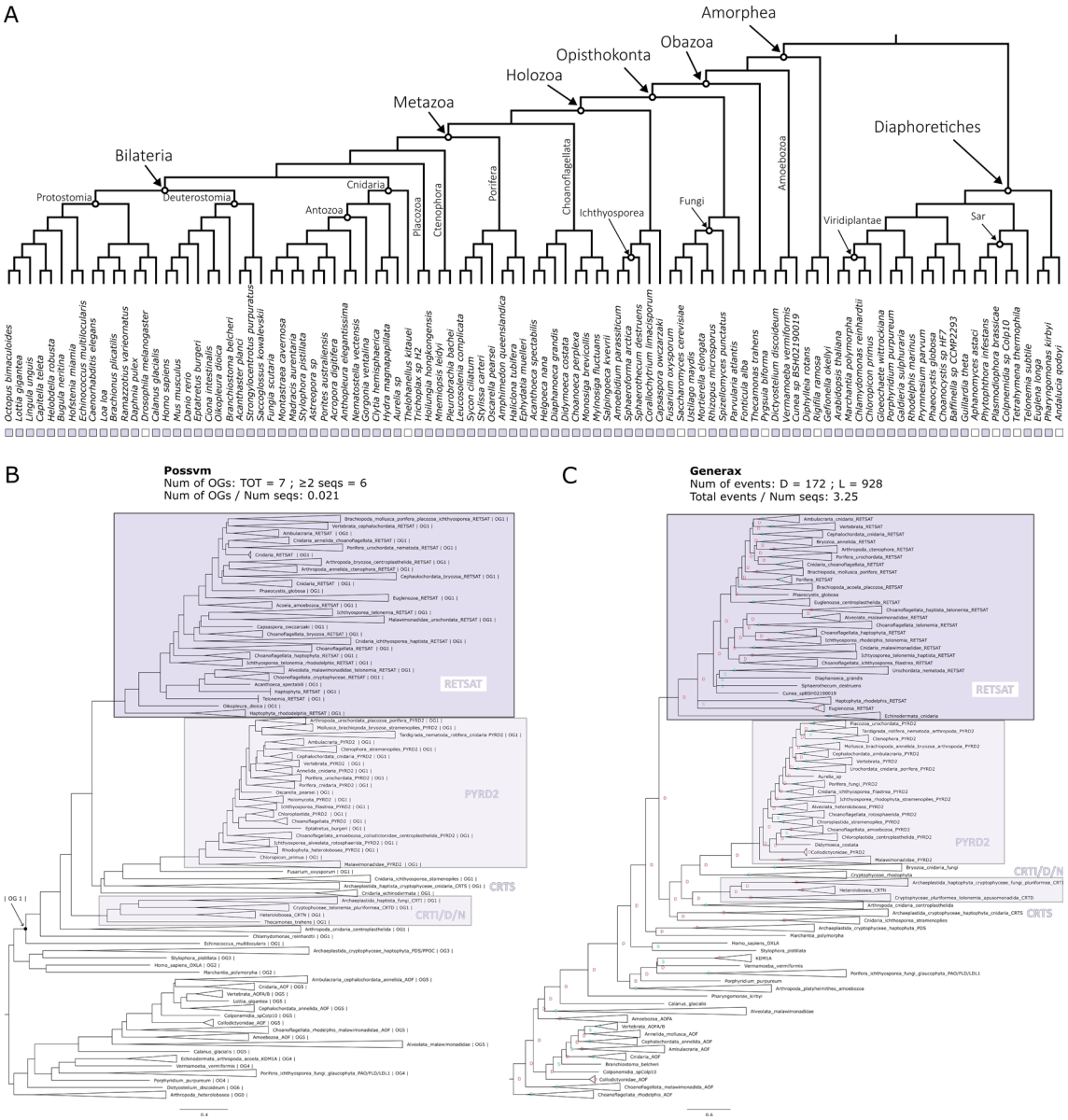


Figure 4.3. Phylogenetic analysis for the RETSAT orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

PNPLA4

The Patatin Like Phospholipase Domain Containing 4 (PNPLA4) enzyme plays a role in the hydrolysis of retinyl esters to retinol (Holmes 2012; Schreiber et al. 2012) (Figure 4.1). It is involved in one other pathway according to KEGG (Table 4.1).

Both OrthoFinder and Broccoli identify one distinct orthogroup for PNPLA4 independent from all other orthogroups (Figure 4.2). The final PNPLA4 orthogroup contains 215 sequences. While being present in both major eukaryotic clades, this orthogroup appears to be missing in some groups of Amorphea, such as the Holomycota branch that includes Fungi (Figure 4.4A).

The phylogenetic analysis clarified the relationship between PNPLA4 and other PNPLA enzymes present in the orthogroup (Figure 4.4 B and C). Possvm identified 9 orthogroups within this family. PNPLA1-5 belonged to the same orthogroup, with PNPLA4 being sister group to the other genes (Figure 4.4B). The GeneRax reconciled tree recovered the same topology and identified a moderate number of events (Figure 4.4C). The phylogenetic analysis also revealed that while the broad PNPLA4 orthogroup included sequences from a wide range of eukaryotic organisms, the PNPLA1-5 sub-clades contained primarily animal sequences. The tight relationship between PNPLA4 and other PNPLA genes is in accordance with evidence suggesting that some of them are also involved in retinol metabolism (Kienesberger et al. 2009; Pingitore and Romeo 2019). Similarly, one cannot rule out the possibility that even more distantly related sequences from non-animal species within the overarching orthogroup might also perform similar functions.

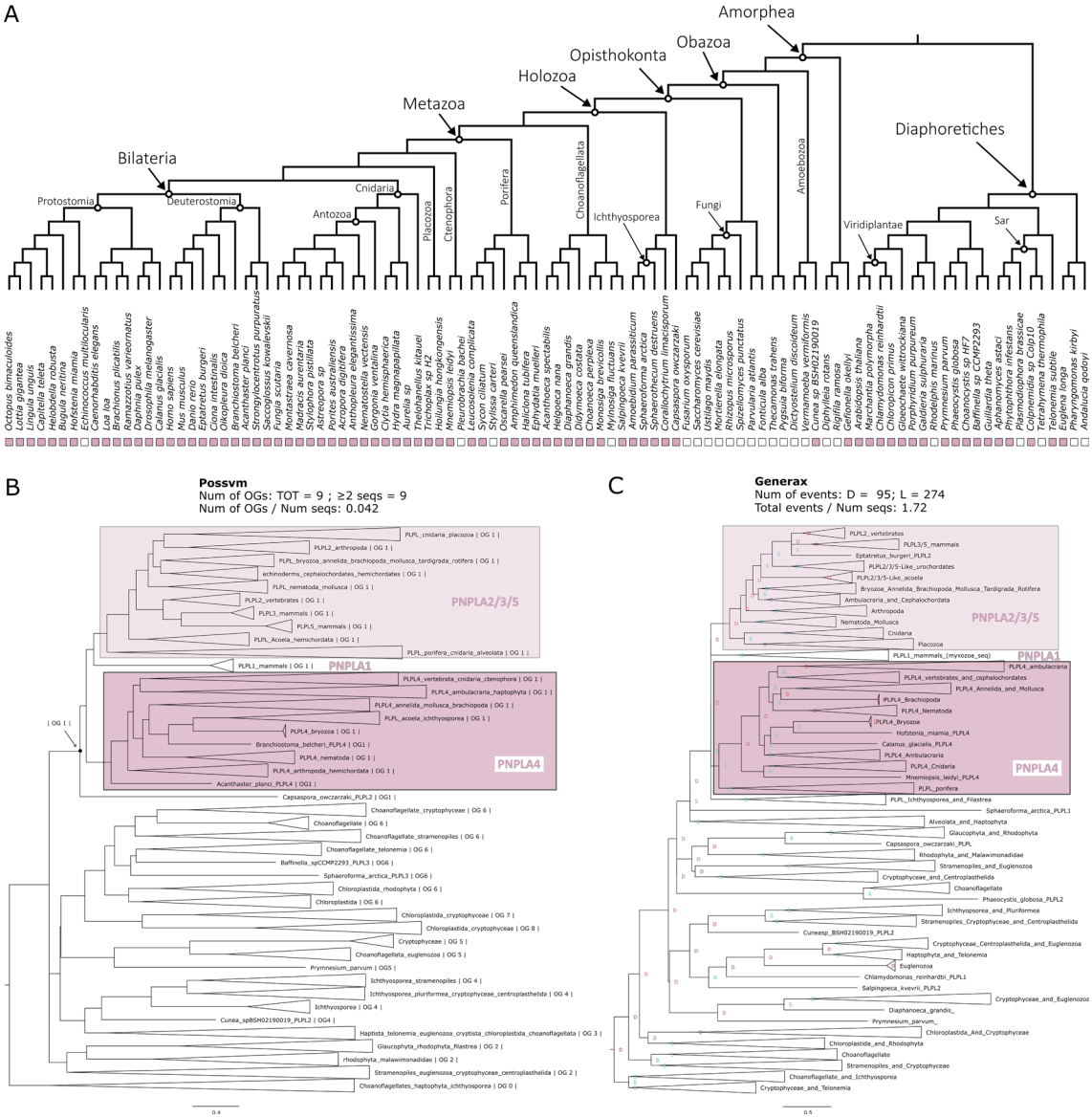


Figure 4.4. Phylogenetic analysis for the PNPLA4 orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

ALDH1

Aldehyde Dehydrogenase 1 Family Member A1 (ALDH1A1 or ALDH1), also known as Retinaldehyde Dehydrogenase 1 (RALDH1), is an enzyme that can catalyse the oxidation of retinal to retinoic acid (or retinoate) (Duester 2000) (Figure 4.1). ALDH1 is involved in two KEGG pathways (Table 4.1).

Both OrthoFinder and Broccoli identify ALDH1 as its own distinct orthogroup (Figure 4.2) and the final merged orthogroup consists of 765 sequences. This orthogroup is ubiquitous, with only a handful of eukaryotic species lacking it (Figure 4.5A).

The phylogenetic analyses revealed a complex substructure within the ALDH1 orthogroup (Figure 4.5 B and C), with Possvm subdividing it into 44 orthogroups, a high number relative to total sequences. ALDH1A, ALDH1B and ALDH2 all coalesce to a same Possvm orthogroup. While the full orthogroup includes other aldehyde dehydrogenases, including ALDH1L, ALDH8A1, ALDH16A1, ALDH9A1 and ALDH5A1. The GeneRax reconciled tree found a very similar topology and identified a relatively high number of evolutionary events (Figure 4.5C). Interestingly, the ALDH1/2 sub-orthogroup predominantly features animal sequences, whereas other ALDH clades encompass a diverse range of eukaryotic species. This suggests a link between the ALDH1/2 expansion within animals and the emergence of vision in these organisms.

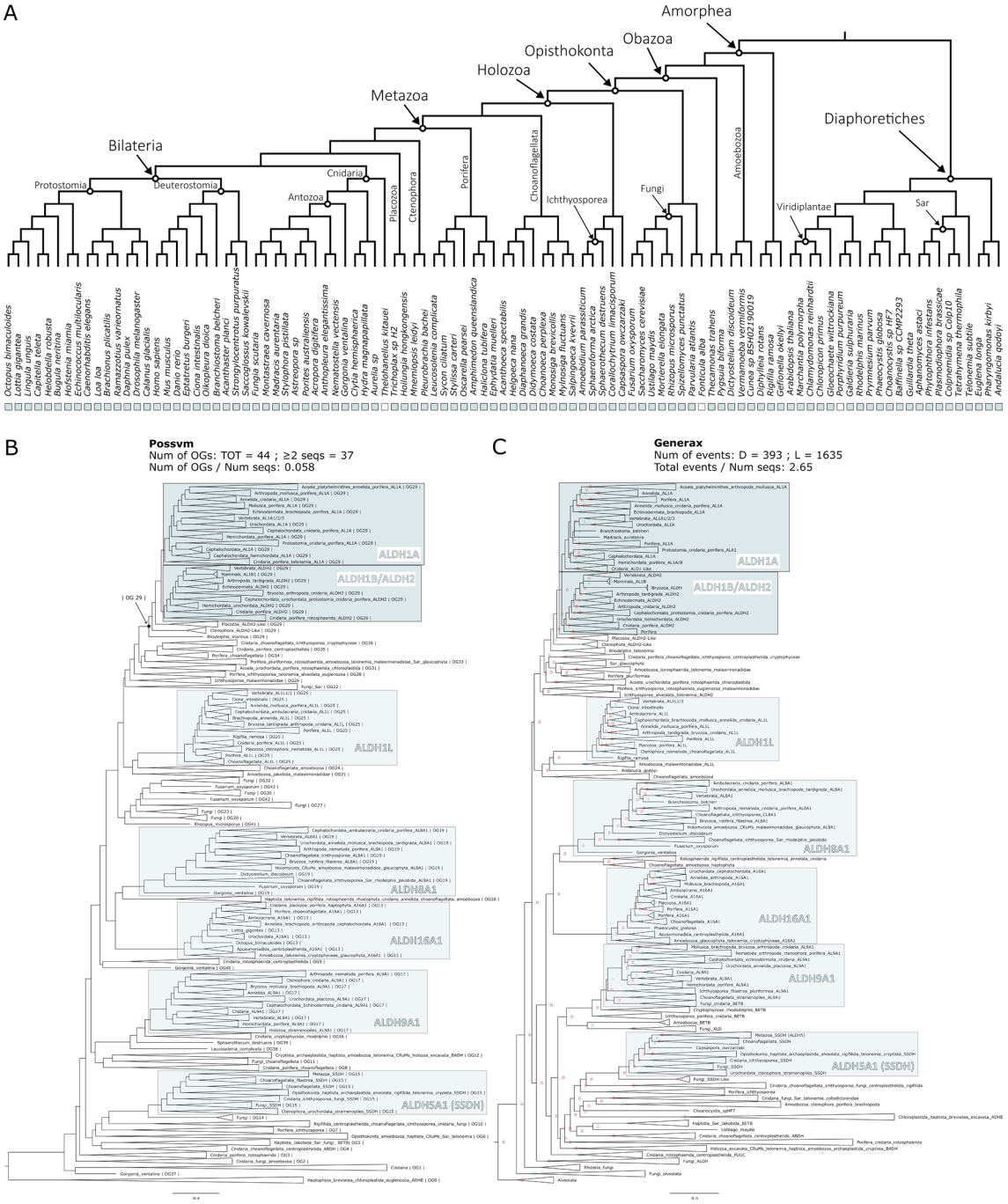


Figure 4.5. Phylogenetic analysis for the ALDH1 orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

BCMO1/RPE65

Beta-carotene 15–15′-monooxygenase (BCMO1), more recently known as Beta-Carotene Oxygenase 1 (BCO1) (Seña et al. 2014), plays a crucial role in converting dietary beta-carotene into retinal by catalysing the symmetric cleavage of beta-carotene to produce two all-trans-retinal molecules (Harrison 2012) (Figure 4.1). Another carotenoid cleavage oxygenase (CCO) enzyme is Retinoid Isomerohydrolase RPE65. RPE65 is expressed in retinal pigment epithelium (RPE) cells where it catalyses the conversion of all-trans-retinyl ester to 11-cis-retinol (Jin et al. 2005; Moiseyev et al. 2005; Redmond et al. 2005). These two essential enzymes are both quite specific to the pathway, with RPE65 being present in only two KEGG pathways and BCMO1 in three (Table 4.1).

BCMO1 and RPE65 are placed in the same orthogroup both by OrthoFinder and by Broccoli (Figure 4.2) confirming that they belong to the same family of enzymes. The complete orthogroup consists in 322 sequences. This orthogroup has a patchy presence throughout most eukaryotic clades (Figure 4.6A).

The phylogenetic analysis for this orthogroup revealed several sub-families. Possvm identified 16 orthogroups within this family, with BCO1, RPE65, as well as BCO2, belonging to the same orthogroup (Figure 4.6B). GeneRax recovers a fairly similar topology and a moderately high number of events (Figure 4.6C). Also in this case, the BCMO1/RPE65 specific subclade appears to be animal-specific; while other subgroups are either widely distributed (like ACOX) or specific to eukaryotic clades distantly related to animals (such as CCD8 and NCED/CCD1).

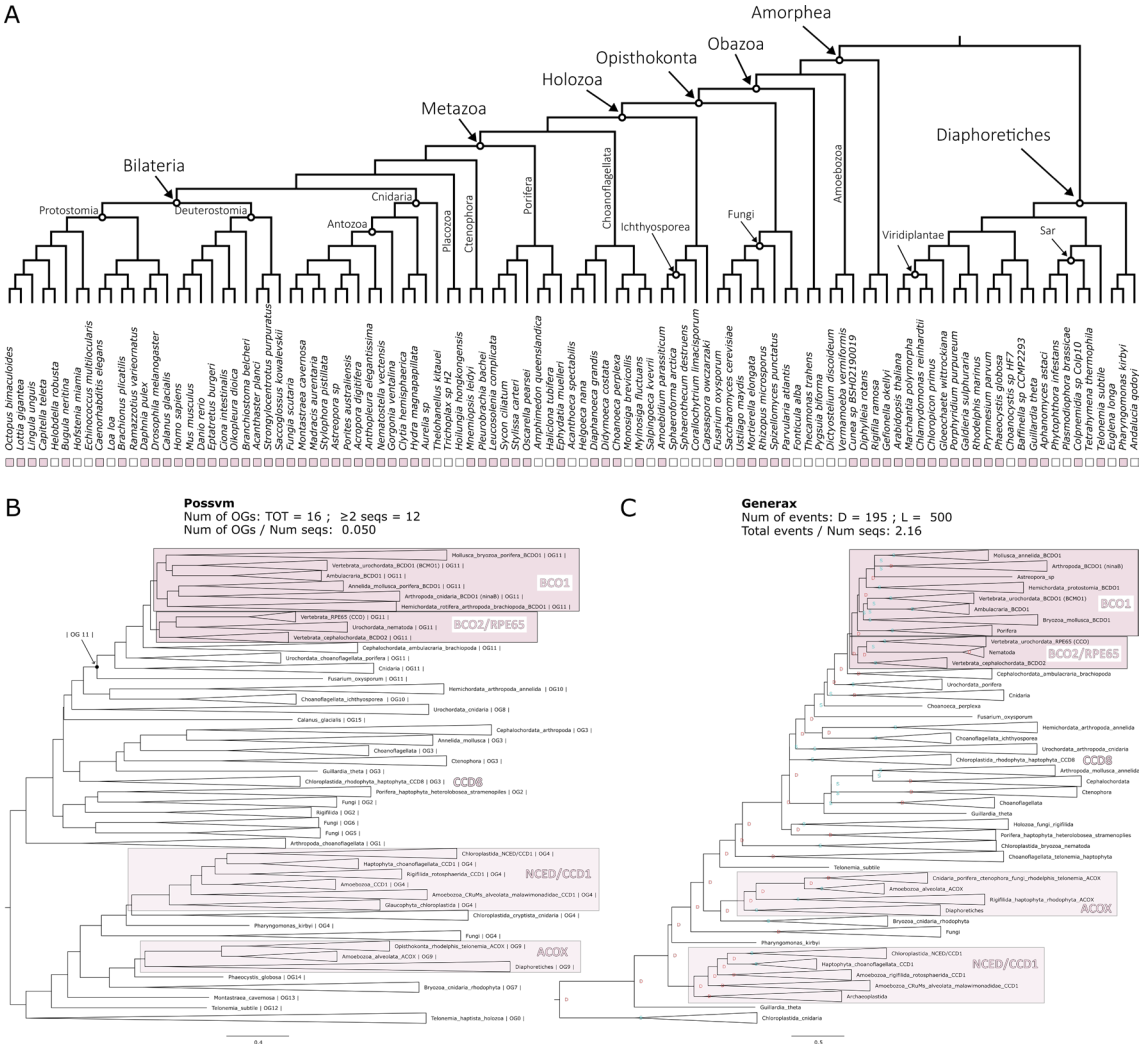


Figure 4.6. Phylogenetic analysis for the BCMO1/RPE65 orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

LRAT

Lecithin Retinol Acyltransferase (LRAT), also known as Phosphatidylcholine--Retinol O-Acyltransferase, catalyses the esterification of all-trans-retinol into all-trans-retinyl ester (Ruiz et al. 1999; Batten et al. 2004) (Figure 4.1). It belongs to three KEGG pathways (Table 4.1).

OrthoFinder and GeneRax orthogroups for this enzyme correspond to each other with high identity (Figure 4.2). The LRAT orthogroup is the smallest, including only 93 sequences. This is reflected in its limited distribution throughout eukaryotes. It is present in most animal clades, with exception of placozoans and ctenophores. However, outside of animals there seems to be very sparse and uneven distribution (Figure 4.7A).

Possvm identifies only 6 orthogroups within LRAT (Figure 4.7B). Interestingly, apart from the orthogroup containing LRAT, there is also an orthogroup containing the related Phospholipase A And Acyltransferase (PLAAT) family of enzymes (Hussain et al. 2017). GeneRax confirms a similar tree topology and identifies a rather high number of events relative to the low number of sequences (Figure 4.7C). The few non-metazoan sequences within the LRAT orthogroup belong neither to the LRAT nor the PLAAT clades in the tree, making it another case in which one of the enzymes most specific to retinol metabolism appears animal specific.

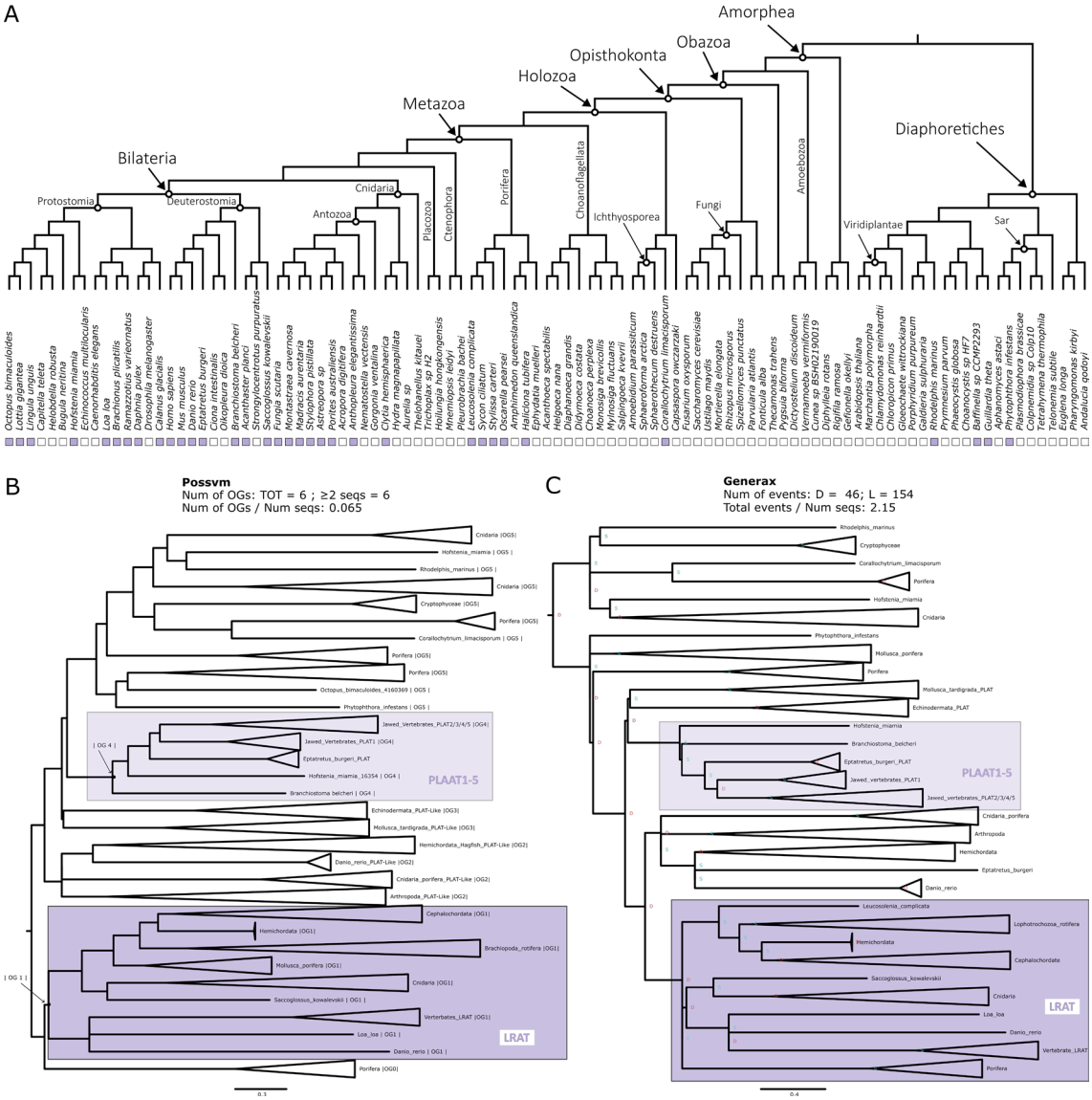


Figure 4.7. Phylogenetic analysis for the LRAT orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

RDH/DHRS

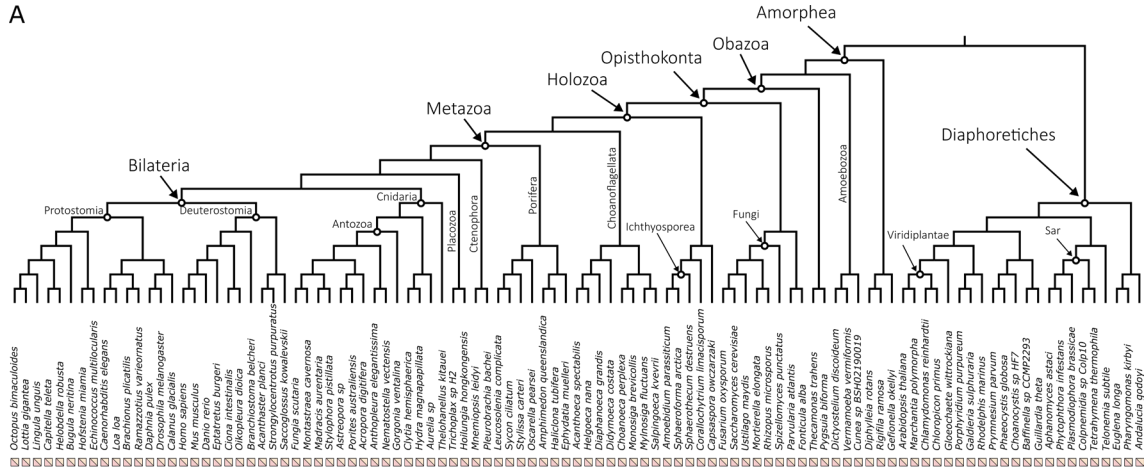
Retinol Dehydrogenase (RDH) enzymes are responsible for the oxidation of retinol to retinal (Sahu and Maeda 2016). RDH5 in particular is responsible for the conversion of 11-cis-retinol to 11-cis-retinal, the visual chromophore (Duester 2000). Other RDHs involved in the retinol metabolism are listed in Table 4.1. These enzymes are quite specific to retinol metabolism, being involved in either two or three KEGG pathways (Table 4.1). RDHs are in turn classified within the broader short-chain dehydrogenases/reductases (SDR) family (Duester 2000; Lhor and Salesse 2014). Other enzymes within this family include members of the Dehydrogenase/Reductases SDR family (DHRS), several of which are also implicated in retinol metabolism (Figure 4.1 and Table 4.1). The DHRS enzymes involved in retinol metabolism belong to a minimum of two up to a maximum of four KEGG pathways (Table 4.1).

The orthogroup analyses reveals a very complex situation for RDH and DHRS enzymes (Figure 4.2). First, there is a substantial difference in results between OrthoFinder that identifies two orthogroups and Broccoli that identifies seven orthogroups containing RDH and DHRS enzymes. Both methods pinpointed two primary orthogroups: one consisting solely of RDH enzymes and another comprising a mix of RDH and DHRS enzymes. Beyond these, Broccoli detected several smaller orthogroups, some leaning towards an RDH profile, while others were more DHRS-specific. Two of the Broccoli orthogroups even share a very small number of sequences with the GUK orthogroup which, being unrelated to the retinol metabolism, was discarded from further analysis. Furthermore, the OrthoFinder DHRS+RDH orthogroup had a small connection with the ADH orthogroup. However, this was negligible (0.09% of identity) and ADH can confidently be regarded as a distinct orthogroup. All these considerations led to the decision to include all RDH and DHRS orthogroups into one big orthogroup for phylogenetic analysis, even when this meant dealing with a large number of sequences. This is in fact the second largest orthogroup examined in this study with a total of 4476 sequences and the only one that is present in every single species examined (Figure 4.8A).

The complexity outlined by the OrthoFinder and Broccoli orthogroup detection is reflected in the complexity of the phylogenetic tree (Figure 4.8 B and C). 207 Possvm orthogroups were defined (Figure 4.8B). The RDH and DHRS enzymes described by KEGG to be involved in retinol metabolism (Table 4.1) are distributed across 6 different possvm orthogroups, which further clade with other members of this expansive family.

GeneRax recovered a largely compatible substructure and revealed a very large number of evolutionary events even for the size of the orthogroup (Figure 4.8C). Overall, not all RDH enzymes belong to a monophyletic clade, and neither do all DHRS enzymes. Instead, monophyletic clades within this broad gene family include enzymes that have been described (based primarily on structure and function) to belong to different sub-families. This underscores the need for a phylogenetic approach to clarify the evolutionary relationships among these enzymes. As mentioned, RDH and DHRS families are part of the extensive SDR super family. Delving deeper into the relationships within other SDR members might shed more light on sub-family connections. However, that would present an extremely challenging task. The current orthogroup (~4500 sequences) already proved to be very demanding between the extensive computational resources needed for certain steps (e.g., GeneRax) and the significant amount of time required to examine the gene trees in detail. Finally, regarding the distribution of specific subgroups, while most subgroups spanned eukaryotes, a handful were animal specific, such as RDH11/12, RDH13, and RDH16/H17B6/DRC7/RDH5/DHRS9. Yet, examining the larger clades these smaller orthogroups are part of reveals the presence of other eukaryotes.

Figure 4.8. Phylogenetic analysis for the RDH/DHRS orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.


A
Amorphea
Obazoa
Opisthokonta
Holozoa
Metazoa
Bilateria
Protostomia
Deuterostomia
Cnidaria
Antozoa
Placozoa
Ctenophora
Porifera
Choanoflagellata
Ichthyosporea
Fungi
Amoebozoa
Diaphoretiches
Viridiplantae
Sar
Octopus bimaculoides
Lottia gigantea
Lingula anatis
Capitella teleta
Helobdella robusta
Bugula nertina
Helobdella mamilla
Strigamia maritima
Caenorhabditis elegans
Loa loa
Brachionus plicatilis
Ramazzottius varieornatus
Daphnia pulex
Drosophila melanogaster
Calanus glacialis
Homo sapiens
Mus musculus
Branchiostoma belcheri
Strongylocentrotus purpuratus
Saccoglossus kowalevskii
Fungia scutaria
Montastraea cavernosa
Nematostella vectensis
Stylophora pistillata
Astreopora sp
Porites australiensis
Acropora digitifera
Hoilungia hongkongensis
Nematostella vectensis
Gorgonia ventalina
Clytia hemisphaerica
Hydra magnapapillata
Mnemiopsis leidyi
Thelohanellus kitauei
Trichoplax sp H2
Hoilungia hongkongensis
Mnemiopsis leidyi
Pleurobrachia bachei
Leucosolenia complicata
Sycon ciliatum
Stylissa carteri
Oscarella pearsei
Salpingoeca rosetta
Monosiga brevicollis
Monosiga fluctuans
Salpingoeca infusionum
Salpingoeca kvevrensis
Salpingoeca urceolata
Sphaeroforma arctica
Sphaeroforma destruens
Corallochytrium limacisporum
Capsaspora owczarzaki
Fonticula alba
Dictyostelium discoideum
Saccharomyces cerevisiae
Ustilago maydis
Mortierella elongata
Rhizopus microsporus
Spizellomyces punctatus
Parvularia atlantis
Fonticula alba
Tetramitus thermophilus
Naegleria gruberi
Trichomonas vaginalis
Dysgusia bilorna
Vermamoeba vermiformis
Cunea sp BSH02190019
Diphylleia rotans
Nyctotherus ovalis
Rigidifila ramosa
Capsellaspora beleziyi
Arabidopsis thaliana
Marchantia polymorpha
Chlamydomonas reinhardtii
Chlorokybus atmophyticus
Chlorella variabilis
Cyanidioschyzon merolae
Porphyridium purpureum
Galdieria sulphuraria
Rhodelphis marinus
Hyrmesium parvum
Phaeocystis globosa
Choanocystis sp HF7
Guillardia theta
Baffinella sp CCMP2293
Aphanomyces astaci
Phytophthora infestans
Plasmodiophora brassicae
Colpnemidia sp Colp10
Colpnemidia sp Colp12
Tetrahymena thermophila
Telonemia subtile
Euglena longa
Phiarymonas kirbyi
Andalucia godoyi
B
Possvm
Num of OGs: TOT = 207 ; ≥2 seqs = 179
Num of OGs / Num seqs: 0.047
DHB4
FABG
DHB8
CBR4
BDHA
DHB14
SDR2/3/4/5
ADRC1/GRDH2
BDH2
PGDH
NOL/NYC1
HCD2
DHRS2/4
TRNH
RHLG
PECR
DECR2
DCXR
SPRE
PTALR
ADRC
HSDL2
RDH14
DHRS13
RDH11/12
RDH13
DHRS12
FEY
WWOX/POR/TIC32
DHRSX
RDH10/RDHE2+DHB11/12+DHRS3
DHR11
DHB12
DHRS7/7B/7C+DHI1
KDSR
RDH8/DHB1
RDH16+H17B6+DR9C7+RDH5+DHRS9
BDH
DHB2/DHI2
C
Generax
Num of events: D = 2391 ; L = 8642
Total events / Num seqs: 2.46
RDH16+H17B6+DR9C7+RDH5+DHRS9
BDH
DHB2/DHI2
RDH8/DHB1
DHRS7/7B/7C+DHI1
KDSR
DHB12
HSDL2
RDH10/RDHE2+DHB11/12+DHRS3
DHR11
RDH14
DHRS13
RDH11/12
RDH13
DHRS12
FEY
WWOX/POR/TIC32
DHRSX
DCXR
SPRE
PTALR
ADRC
DHRS2/4
TRNH
RHLG
PECR
DECR2
DHB4
FABG
DHB8
CBR4
BDHA
DHB14
SDR2/3/4/5
PGDH
NOL/NYC1
HCD2
BDH2
ADRC1/GRDH2

DGAT1

Diacylglycerol O-Acyltransferase 1 (DGAT1) is known primarily for its role in triacylglycerol synthesis (Bhatt-Wessel et al. 2018). However, it has also been implicated in the retinol metabolism as an alternative to LRAT in the esterification of retinol to retinyl esters (Orland et al. 2005) (Figure 4.1). DGAT1 is involved in four metabolic pathways according to KEGG (Table 4.1).

KEGG places both DGAT1 and DGAT2L4 (see below) within the same generic "DGAT" node in the pathway (Figure 4.1 and Table 4.1). While they are involved in catalysing the same reaction, the orthogroup detection analysis clearly indicates that DGAT1 and DGAT2L4 are independent orthogroups, with both OrthoFinder and Broccoli keeping them separate (Figure 4.2). Therefore, the phylogenetic analysis was performed separately for these two orthogroups. The DGAT1 orthogroup contains 246 sequences and appears to be present throughout all eukaryotes with only a handful of species missing it (Figure 4.9A).

The Possvm analyses revealed a relatively simple substructure with only 7 orthogroups (Figure 4.9B). DGAT1 itself is monophyletic and belonging to one orthogroup. The Sterol O-Acyltransferase (SOAT) family appears to be closely related to DGAT1. The same substructure was described by GeneRax that also revealed a relatively low number of evolutionary events within this orthogroup (Figure 4.9C). The DGAT1 sub-orthogroup defined by Possvm includes sequences from across eukaryotes.

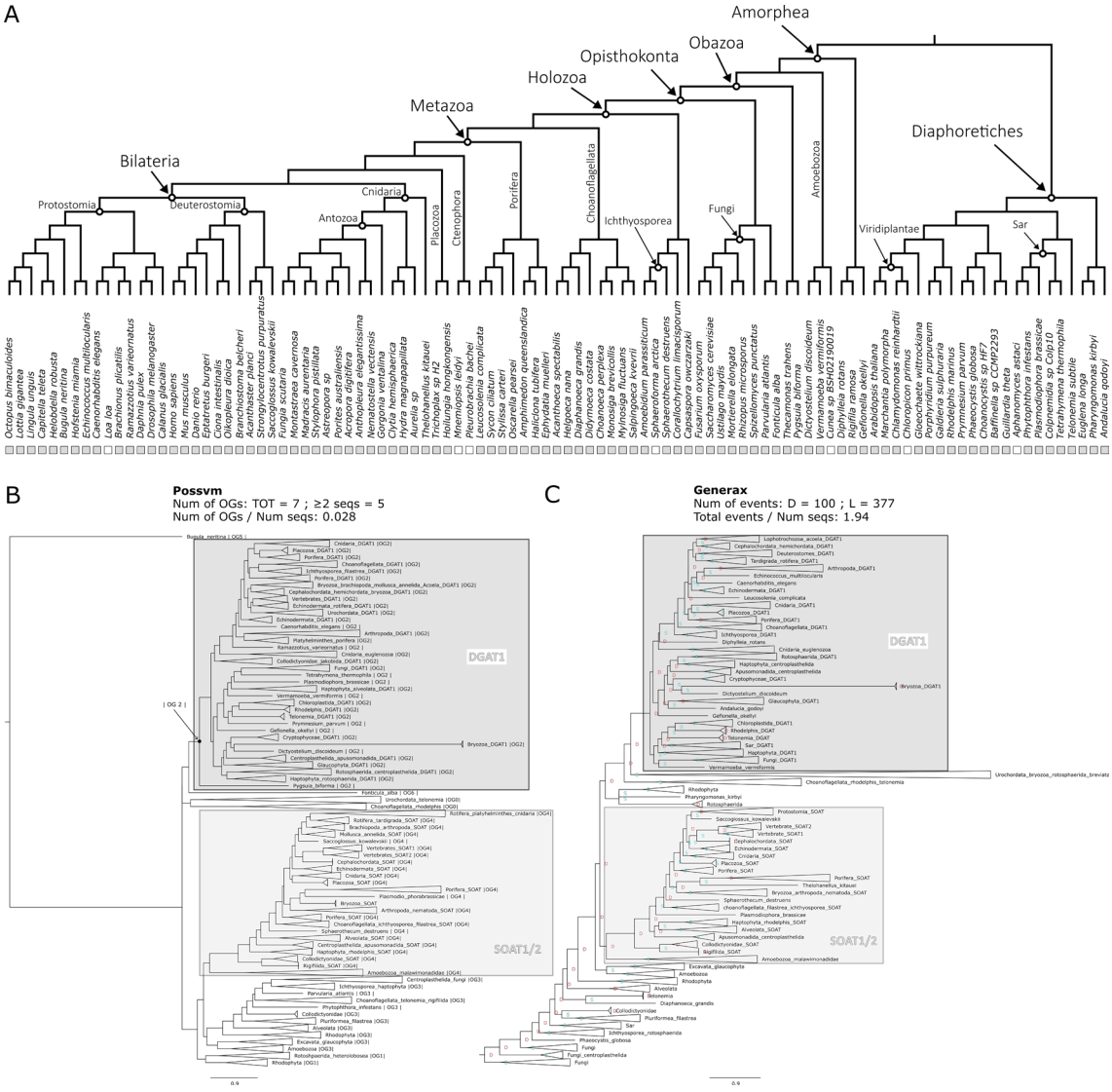


Figure 4.9. Phylogenetic analysis for the DGAT1 orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

DGAT2LA4

Diacylglycerol O-Acyltransferase 2-Like Protein 4 (DGAT2L4), also known as Acyl-CoA Wax Alcohol Acyltransferase 2 (AWAT2), is primarily known for its role in the production of wax esters (Cheng and Russell 2004). It has also been recently implicated in the conversion of retinol to retinyl ester (Kaylor et al. 2014; Arne et al. 2017; Blaner 2017) (Figure 4.1). According to KEGG this enzyme is involved in three metabolic pathways (Table 4.1).

Although DGAT2LA4 indeed seems to be involved in the same step as DGAT1 (and LRAT), it appears to form its own distinct orthogroup (see above) (Figure 4.2). This orthogroup includes 372 sequences and is present in all eukaryotes with few species missing it (Figure 4.10A).

Possvm identified 23 orthogroups, which quite high for the number of sequences (Figure 4.10B). DGAT2L4 forms a monophyletic clade with DGAT2L2, DGAT2L3, DGAT2L6 and DGAT2. While DGAT2L1 and DGAT2L5 form another monophyletic clade, sister group to the previous one (Figure 4.10B). Both clades, together with other less well characterized sequences, belong to one Possvm orthogroup. The same relationships are maintained in the reconciled tree by GeneRax that calculated quite a high number of events (Figure 4.10C). While the clades encompassing the DGAT2L and DGAT2 genes are specific to animals, the same Possvm orthogroup contains various non-metazoan sequences. This implies that this gene family existed anciently, even if animal-specific expansions gave rise to the recognized enzymes with a marginal role in retinol metabolism.

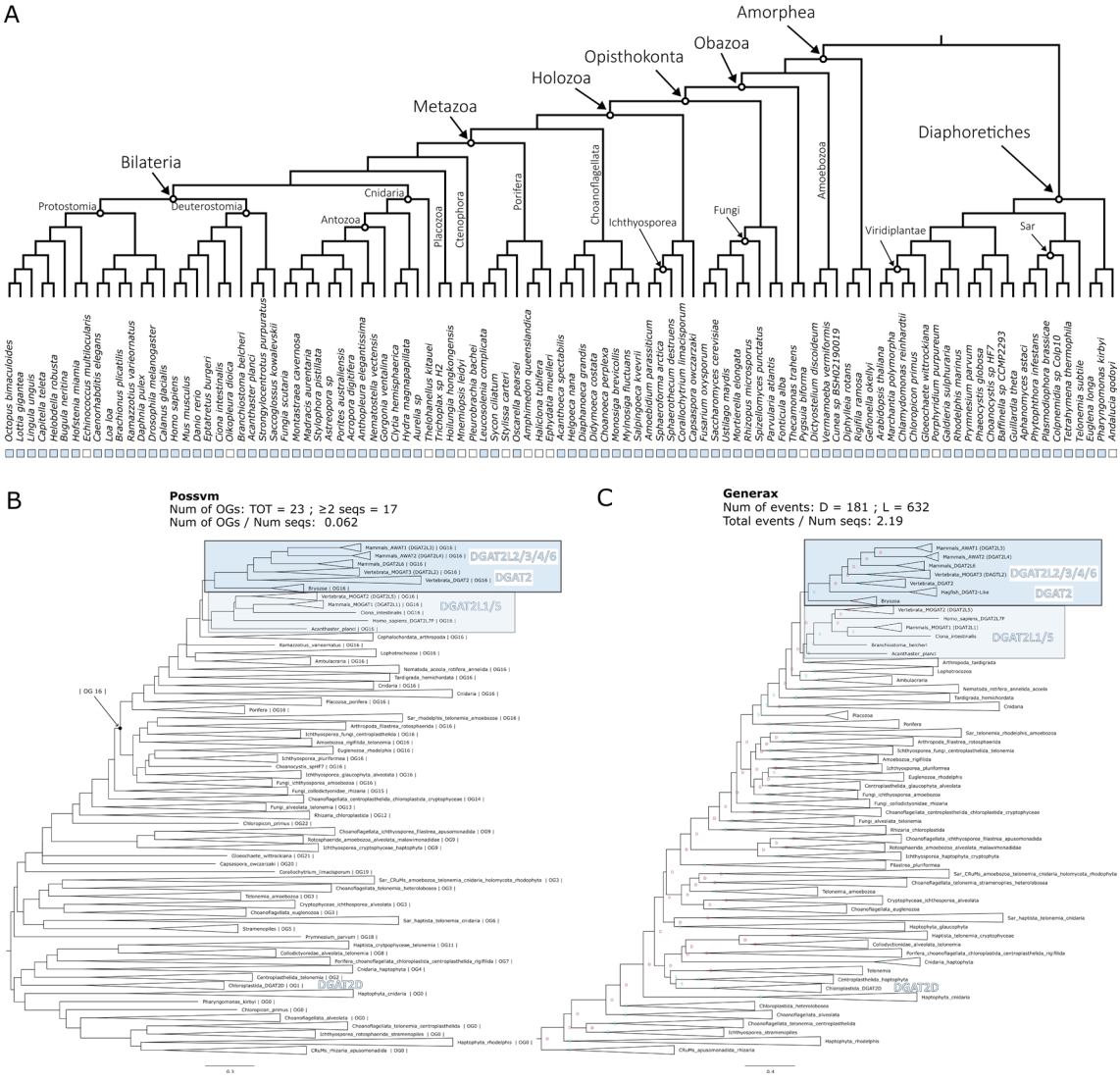


Figure 4.10. Phylogenetic analysis for the DGAT2L4 orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

CYP

Cytochrome P450 (CYP) enzymes represent a large and diverse family of heme-containing enzymes involved in the synthesis and metabolism of a wide range of compounds (Zhao et al. 2021). The number of CYP enzymes is so vast that it is generally considered to be a super family in turn subdivided into families and sub-families (Nelson 2018). For example, the CYP27C1 enzyme, the most specific to the retinol metabolism (Table 4.1), belongs to the family 27, sub-family C, and is the member 1. It catalyses the 3,4 desaturation of all-trans-retinol to all-trans-3,4-didehydroretinol (Enright et al. 2015; Kramlinger et al. 2016; Corbo 2021) (Figure 4.1). The other CYP enzymes involved in the retinol metabolism have varied degree of specificity and are listed in Table 4.1.

While being a vast family, the orthogroup identification was straightforward, with OrthoFinder and Broccoli results coinciding (Figure 4.2). The total orthogroup contained 4499 sequences, making it the largest group examined in this study. The distribution also spans all of Eukarya with only three species of the 101 examined lacking it (Figure 4.11A).

Possvm identified 74 orthogroups (Figure 4.11B), meaning that while being slightly larger than the RDH/DHRS orthogroup, it is overall much less fragmented. Nevertheless, the CYP enzymes described to be involved in the retinol metabolism (Table 4.1) are not all belonging to the same Possvm orthogroup, nor to one monophyletic clade, but rather span 5 separate monophyletic clades. These groups are confirmed with the GeneRax reconciliation (Figure 4.11C) that also identifies a relatively low amount of duplication and loss events considering the number of sequences in the orthogroup. Overall, monophyletic clades encompassing CYP enzymes implicated in retinol metabolism contain sequences spanning most eukaryotic groups.

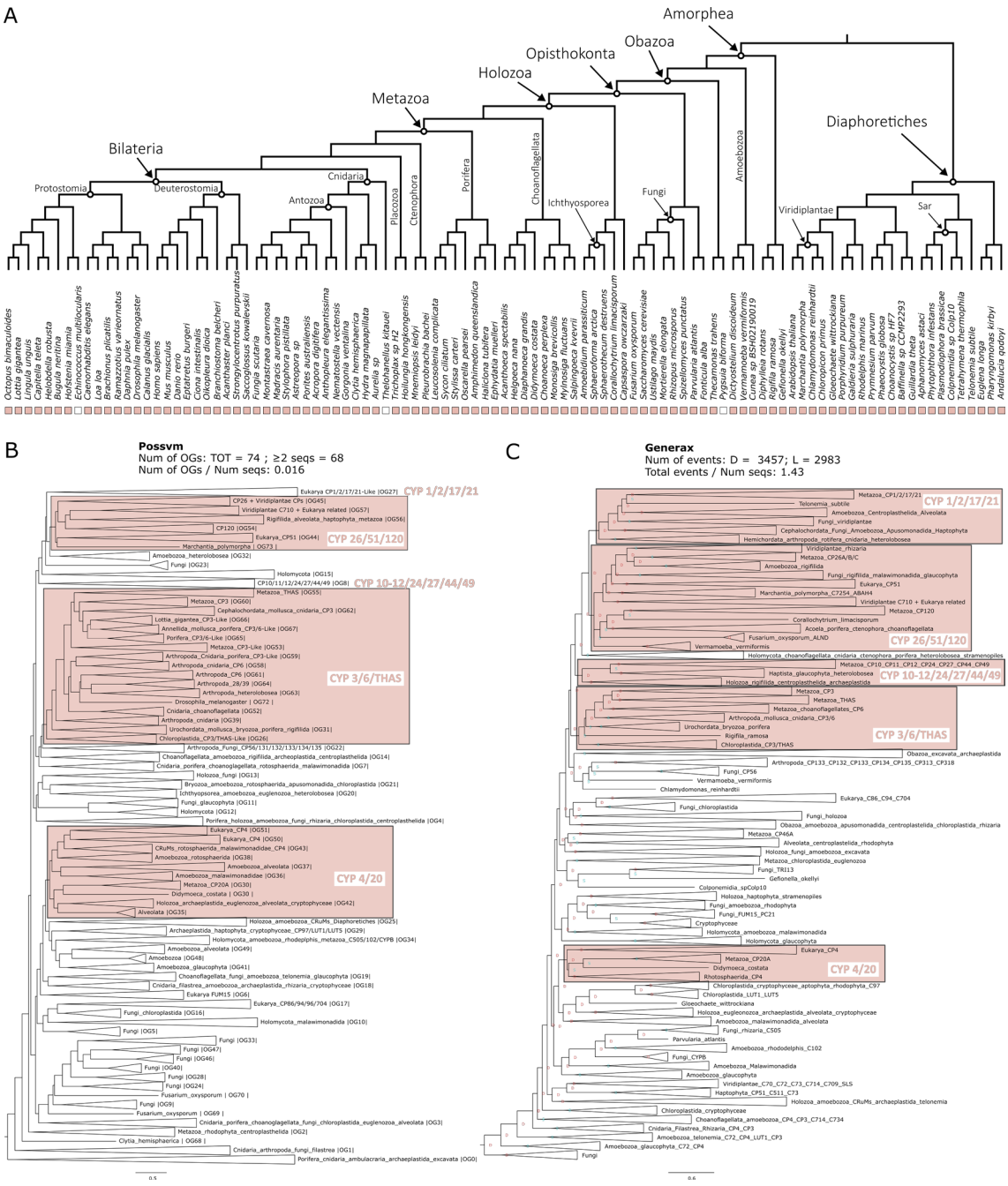


Figure 4.11. Phylogenetic analysis for the CYP orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

AOX

Aldehyde Oxidase 1 (AOX1) is responsible for the oxidation of a wide variety of aldehydes to their corresponding carboxylic acids (Terao et al. 2016). Within the retinol metabolism it can oxidise retinal to retinoate (Terao et al. 2016) (Figure 4.1), although the primary enzyme for this is ALDH1 (see above). Overall AOX1 is not to be considered specific to the retinol metabolism (Table 4.1).

The identification of the AOX orthogroup presented slight differences between OrthoFinder, which found one orthogroup, and Broccoli, that split the family into two orthogroups, the AOX and the AAO (Abscisic-aldehyde oxidase), a group of aldehyde oxidases primarily known in plants (Seo et al. 2000). The total orthogroup of AOX includes 599 sequences. It is overall present in all eukaryotes with some exceptions, e.g., ctenophores (Figure 4.12A).

Possvm identified 25 orthogroups (Figure 4.12B). The phylogenetic analysis uncovered how the Xanthine Dehydrogenase (XDH) family is closely related to the AOX. While the AAO (present primarily in Diaphoretiches) is more distantly related. This is confirmed in the reconciled GeneRax tree that also revealed a moderate number of events (Figure 4.12C). Interestingly, while the AOX clade is limited to a specific subset of animal species, the closely related XDH clade encompasses sequences from a diverse array of eukaryotes.

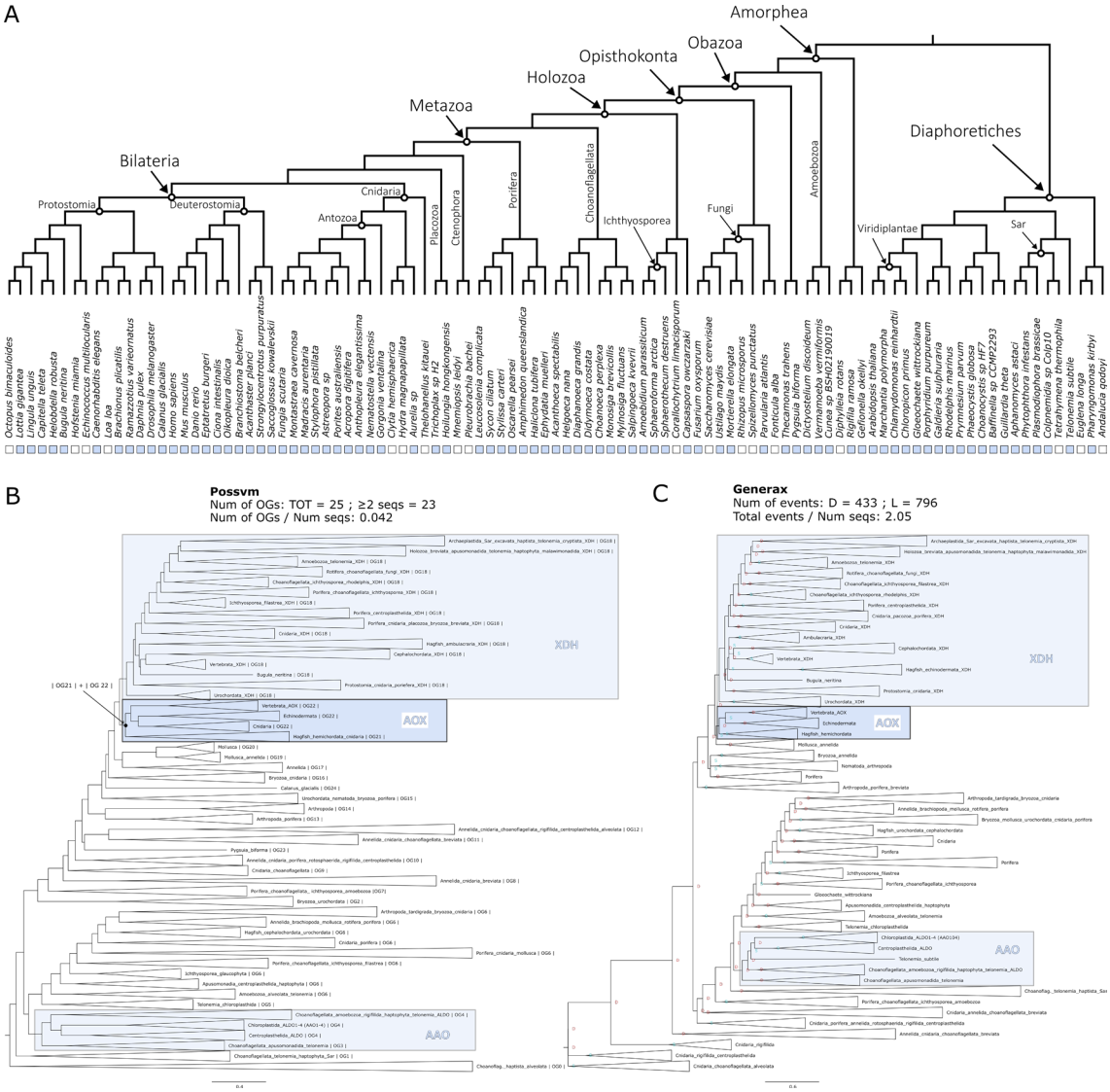


Figure 4.12. Phylogenetic analysis for the AOX orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

ADH

Alcohol dehydrogenase (ADH) enzymes play crucial roles in the metabolism of alcohols (Edenberg 2007). In the retinol metabolism ADHs play a role in the oxidation of retinol to retinal (Duester 2000) (Figure 4.1). Although RDH is the primary enzyme for this reaction, particularly within the retina, ADHs can also contribute and within humans are especially used in non-visual related tissues such as the liver (Duester 2000). Seeing as ADHs are involved in metabolising a wide variety of alcohols it is not surprising that they are involved in numerous other pathways other than the retinol metabolism (Table 4.1).

During the identification of an orthogroup for ADH, OrthoFinder placed all sequences in one orthogroup, while Broccoli split the family into two orthogroups. One primarily comprised ADH sequences, while the other was a mixed group that incorporated the related Sorbitol Dehydrogenase (SORD) (Figure 4.2). The merged orthogroup consisted of 955 sequences and was present in all but one species (Figure 4.13A).

Ortholog analysis with Possvm revealed a complex substructure, with 59 orthogroups identified (one of the highest numbers relative to orthogroup size) (Figure 4.13B). Possvm split the various ADH enzymes into different orthogroups, with ADH5 being the most distantly related. Nevertheless, all ADHs belonged to a larger monophyletic group. Other families picked up in this broad orthogroup are Cinnamyl alcohol dehydrogenase (CADH), Succinate-semialdehyde dehydrogenase (SUCD), and Sorbitol Dehydrogenase (SORD). The GeneRax reconciled tree maintains the same overall topology and a large number of events were calculated, one of the highest relative to number of sequences (Figure 4.13C). The ADH1/4/6/7 group seems to represent a mammalian-specific expansion within the family. In contrast, ADH5 appears ancient, comprising sequences from many different eukaryotic groups.

Figure 4.13. Phylogenetic analysis for the ADH orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

UGT

UDP-glucuronosyltransferase (UGT) enzymes are involved in the process of glucuronidation of small lipophilic molecules, whereby a glucuronic acid is transferred from a UDP-glucuronic acid to the small molecule, making it more water soluble and therefore easier to excrete from the body (Rowland et al. 2013). In mammals there are four UGT families: UGT1; UGT2; UGT3; and UGT8 (Meech et al. 2019). UGTs are involved in the regulation of retinoid levels in the body; by glucuronidating all-trans-retinoate to all-trans-retinoyl beta-glucuronide it facilitates the excretion of this molecule (Meech et al. 2019) (Figure 4.1). Overall, this enzyme family is very broad spectrum (Table 4.1) and involved only marginally in the retinol metabolism, nevertheless we included it in our evolutionary study.

UGTs are clearly identified as being an independent orthogroup by both OrthoFinder and Broccoli (Figure 4.2). This orthogroup consists of many sequences (1005 sequences). Interestingly, while present in both major branches of eukaryotes, it appears to be missing in several clades, including several unicellular holozoans (such as ichthyosporeans) that are closely related to animals, although it is present in the sister group to animals, the choanoflagellates (Figure 4.14A).

The phylogenetic analysis uncovers that UGT1 and UGT2 are closely related to each other, as are UGT3 and UGT8. However, all of them belong to a single monophyletic clade, which Possvm identifies as one orthogroup (Figure 4.13B). The GeneRax reconciled tree maintains this topology (Figure 4.14C). Overall, Possvm identifies a total of 21 orthogroups and GeneRax identifies the lowest ratio of events to sequences from all orthogroups examined. Collectively, this indicates that the UGT orthogroup is rather conserved. The UGT1/2/3/8 monophyletic clade predominantly consists of deuterostome (vertebrates and their close relatives) sequences within a Possvm orthogroup that includes only animal sequences. Nevertheless, the rest of the broad orthogroup contains a diverse array of eukaryotic sequences, including an apparently plant specific clade of UGTs.

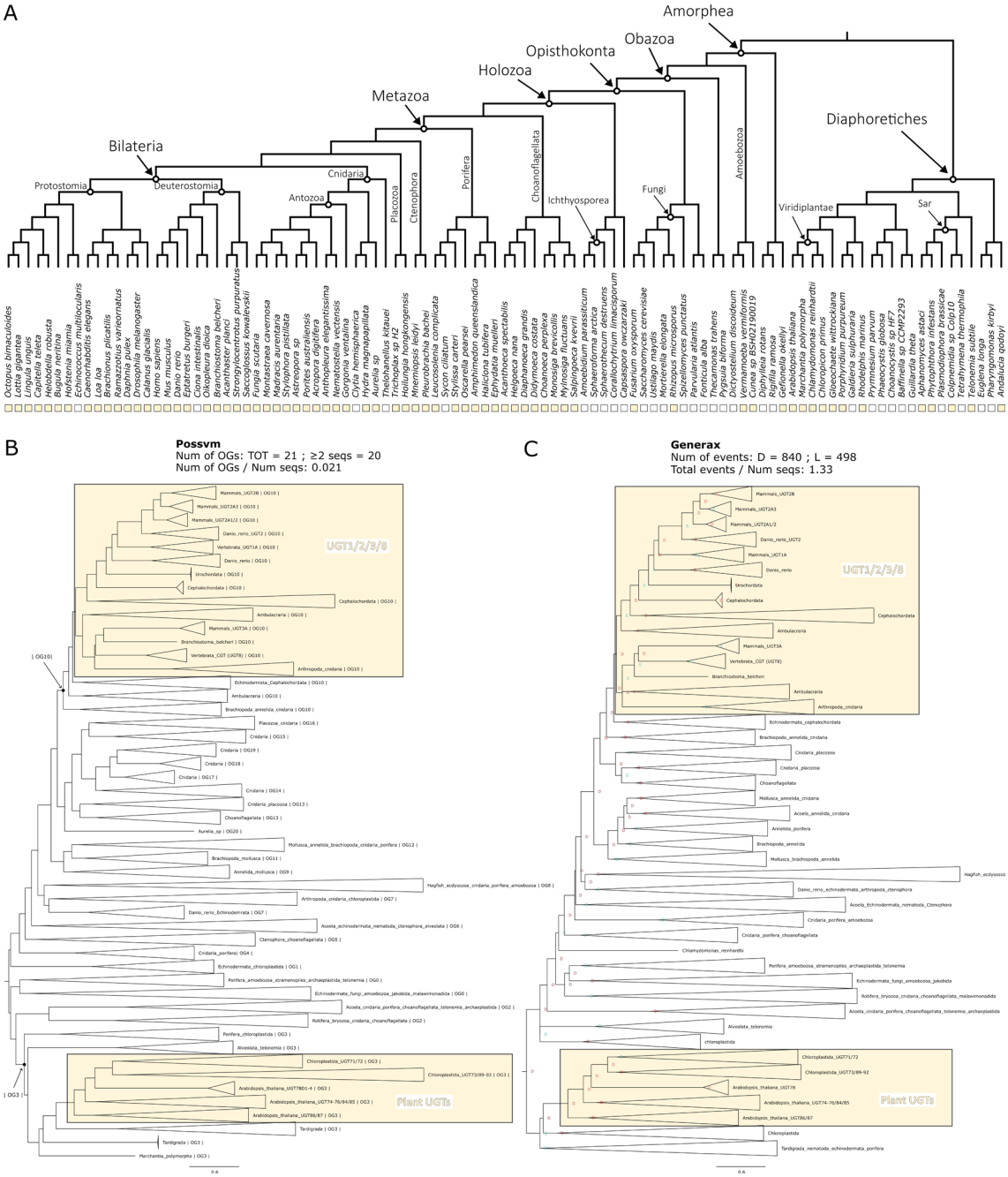


Figure 4.13. Phylogenetic analysis for the UGT orthogroup. (A) Distribution of the orthogroup throughout Eukarya. **(B)** Phylogenetic tree analysed with Possvm for identification of sub-orthogroups. Number of Possvm orthogroups identified (Total, and with ≥ 2 sequences) are provided, as well as the ratio between number of Possvm sub-orthogroups per number of sequences in full orthogroup are indicated. **(C)** Gene tree to species tree reconciliation using GeneRax highlighting speciation and duplication events at each node. Numbers of duplications (D) and losses (L) are provided, as well as the ratio of total events (D+L) per number of sequences in the orthogroup.

Conclusions

Vision, a distinguishing feature of the animal kingdom, hinges on a specific light-sensitive molecule initiating the phototransduction pathway. This molecule is the visual chromophore 11-cis-retinal bound to the membrane protein opsin in photoreceptor cells. When 11-cis-retinal absorbs light, it isomerises into all-trans-retinal, setting off the phototransduction process (Lamb 2020). Continuous light detection demands that this chromophore be perpetually restored to its original 11-cis state. This is obtained through the retinol metabolism, a pathway essential to both vision and other biological functions (Blomhoff and Blomhoff 2006). Thus, understanding the origins and evolution of vision necessitates exploring the evolution of retinol metabolism, which sustains our light sensitivity.

The aim of this chapter was to comprehensively explore the evolution of retinol metabolism genes to provide insights on the origin of the pathway. Initially, I explored the relationships among the enzymes integral to the pathway, grouping them into overarching gene families or orthogroups. Subsequently, I analysed the distribution of these orthogroups across eukaryotes. Finally, I outlined the evolutionary events that have shaped the history of each respective orthogroup.

Enzymes can be classified into families based on several criteria, with enzymatic activity being one of the most predominant (Webb 1992). Through my orthogroup analyses, I sought to determine if the enzymes involved in retinol metabolism could be grouped according to their evolutionary relationships and how these groups align with established enzymatic families. My results suggest that enzymes integral to retinol metabolism can be categorized into 12 distinct orthogroups (Figure 4.2, Table 4.3). Some of these enzymes play pivotal roles in the critical steps for recycling 11-cis-retinal, while others have more peripheral functions (Figure 4.15A). This analysis generally aligned with the established enzymatic families, yet it also shed light on some unexpected findings. For instance, while Diacylglycerol O-Acyltransferase enzymes have conventionally been classified under a single overarching family, my findings provide compelling evidence that DGAT1 exhibits significant evolutionary divergence from DGAT2 and its related molecules, such as DGAT2L4, warranting their categorization into distinct orthogroups

(Figure 4.2). Another example involves the enzymes responsible for converting retinol to retinal, which belong to sub-families within the expansive SDR family. While prevailing nomenclatures suggest a distinction between RDH and DHRS, the orthogroup analyses suggested a more complex web of relationships (Figure 4.2), subsequently corroborated by phylogenetic analyses (Figure 4.8). Ultimately all this suggests that within the SDR family, phylogenetic relationships may define different sub-families compared to the currently established ones.

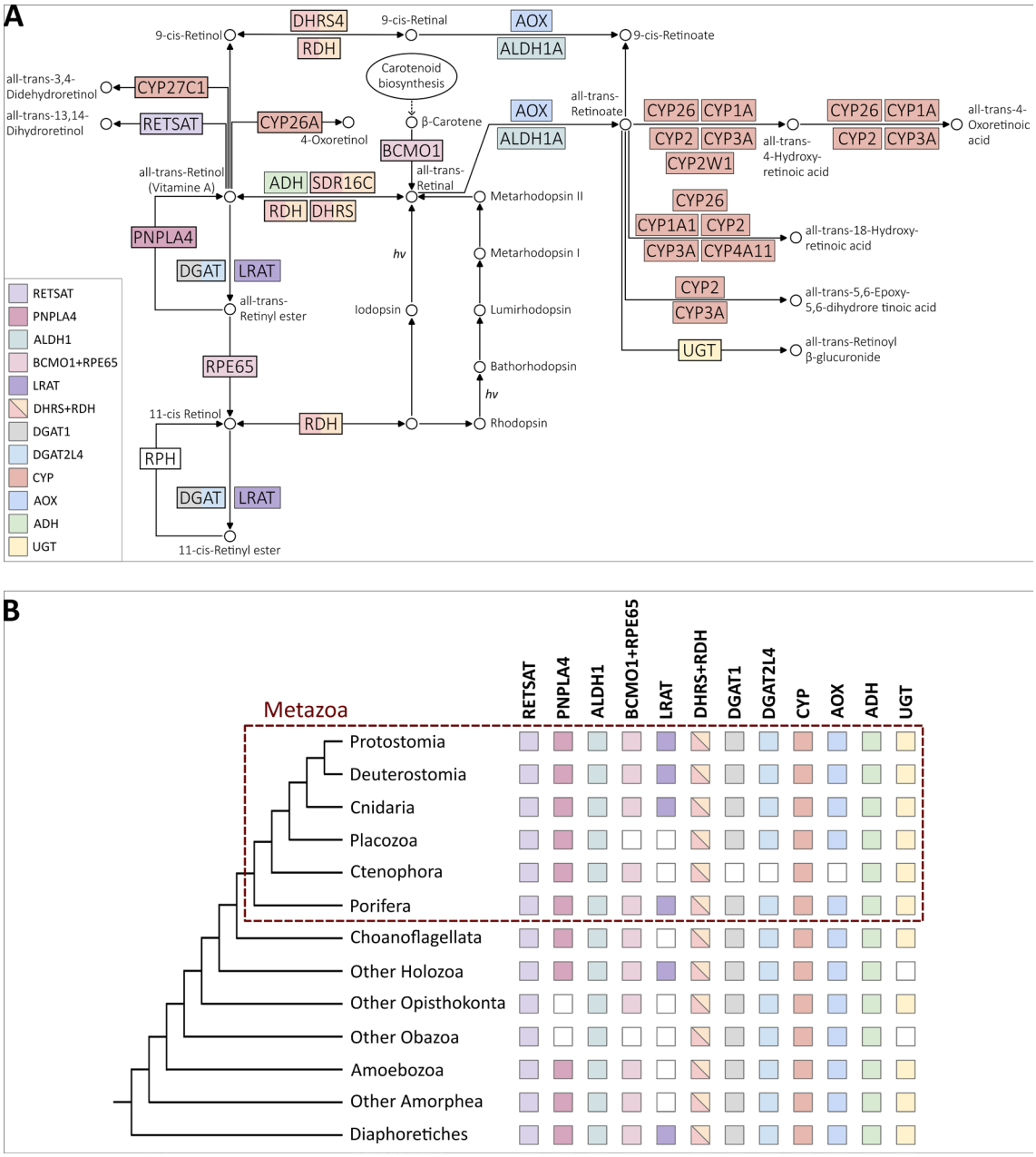


Figure 4.15. Summary of the orthogroups to which enzymes involved in retinol metabolism belong to and their distribution. (A) Overlap of the retinol metabolism related orthogroups identified in this study with the KEGG map00830 pathway components. Some enzymes belong to the same orthogroup (e.g., RDH

and DHRS) while some are split into different orthogroups (e.g., DGAT1 and DGAT2L4). **(B)** Orthogroups are ancient and widely distributed throughout Eukarya, except LRAT.

Examining the distribution of orthogroups throughout eukaryotes, the first consideration is that most orthogroups have ancient origins, spanning the major eukaryotic clades (Figure 4.15B). The only exception is LRAT that appears to be present primarily in animals (except placozoans and ctenophores) and in a handful of other species. While orthogroups tend to be present in all major eukaryotic clades, only one orthogroup (RDH/DHRS) was present in every single species examined (Figure 4.8). Within animals, only the two non-bilaterian phyla of placozoans and ctenophores seem to lack some key orthogroups, notably BCMO1/RPE65 and LRAT, which are integral to the retinol metabolism pathway. This observation is intriguing for two main reasons. First, the absence of these enzymes might suggest that these two phyla have evolved a variant of the retinol metabolism where these missing enzymes are replaced by others. For instance, the function of LRAT can also be carried out by DGAT1 and DGAT2L4. Secondly, even though sponges lack opsins, they possess all the retinol metabolism orthogroups (although not every sponge species has all orthogroups). This has interesting implications in understanding the origin of vision. It reinforces the growing interest surrounding potential opsin-independent light-sensing processes in sponges (Wong et al. 2022), as discussed in Chapter 3.

Further insights came with the detailed phylogenetic analyses. The approach of using both Possvm and GeneRax provided a dual advantage: Possvm facilitated the swift identification and annotation of sub-orthogroups; while GeneRax ensured a thorough and accurate description of evolutionary events of duplication, speciation, and losses. These events serve as a gauge of the evolutionary intricacies (as opposed to a more linear progression) and can also account for the degree of sub-family diversity observed within orthogroups. These analyses facilitated the identification of orthogroups with particularly complex evolutionary histories. For instance, the ADH orthogroup showcased one of the highest ratios of sub-orthogroups to sequences and also had one of the highest ratios of evolutionary events to sequences. Furthermore, through these analyses, we can juxtapose orthogroups of comparable sizes. This allows, for example, to discriminate between large orthogroups with simple subgroup structure, such as CYP, and large orthogroups with more intricate substructures, like RDH/DHRS.

Another compelling insight from the phylogenetic analyses was the refined understanding of distribution patterns within Eukarya, allowing for a detailed mapping of subgroup distributions beyond just the overarching orthogroups. An intriguing observation that emerged was that precisely the enzymes that are most specific to retinol metabolism (PNPALA4, ALDH1, BCMO1/RPE65) presented a distinctive pattern where the overarching orthogroups spanned across Eukarya, but the sub-families actively utilized in the pathway were predominantly animal specific. This suggests that animal specific expansions of these enzyme families may have evolved concurrently with the development of vision. However, it should be noted that other sub-families within those orthogroups might still hold the potential to execute similar functions.

To conclude, the ancient origin of the enzyme families involved in the replenishment of 11-cis-retinal suggest that the foundational molecular framework was already in place well before the emergence of vision. Expansions in these enzymes families during early animal evolution could have steered this pathway towards its specialized role in visual functions. This study's detailed reconstruction of retinol metabolism enzyme families paves the way for exploring a whole new set of questions regarding the evolution of vision. One vital question that arises is whether the enzymes identified bioinformatically within early-branching animals, truly operate in a physiological setting to facilitate the recycling of 11-cis-retinal in these animals. This is an especially fascinating enquiry to pursue for organisms like sponges which, lacking opsins, in theory should not rely on this pathway for vision. Another intriguing avenue for future research is to understand the evolution of the cell types involved in this pathway. In the human retina, for instance, several cell types, including the Retinal Pigment Epithelium (RPE) cells and Mueller cells, play roles in the visual chromophore replenishment (Arshavsky 2002; Mata et al. 2002; Thompson and Gal 2003; Moiseyev et al. 2005; Strauss 2005). It would be interesting to determine whether homologous functions are carried out by homologous cell types throughout animals, including in early-branching animals. Alternatively, different animals might manage parts of the pathway using unrelated cell types or execute the functions within a single cell type. Thus, future research on the evolution of vision should expand their focus beyond just the evolution of photoreceptor cells, which perform phototransduction, to also include cell types implicated in retinol metabolism.

Methods

Identification of orthogroups for retinol metabolism enzymes.

Species list and species tree

To understand the evolution of the retinol metabolism, I selected 101 eukaryotic species (Table 4.2 and Extended Table 4.2) in which to search for the genes involved in the pathway. The choice of species was based on a combination of balanced taxonomic sampling throughout Eukarya and quality of the proteomes. The latter was assessed using BUSCO (v4.0.6) (Simão et al. 2015; Waterhouse et al. 2018) with the eukaryota_odb10 database. The final selection included 50 animals, of which 25 non-bilaterians, 13 unicellular holozoans closely related to animals, and various other species from all major eukaryotic clades.

The single-copy BUSCO genes obtained from the BUSCO analysis were also used to construct a species tree. This is because knowledge of species relationships can be used both for orthogroup inference with the OrthoFinder software (Emms and Kelly 2015; Emms and Kelly 2019) and to construct species-tree-aware gene trees (Boussau and Scornavacca 2020) using software such as GeneRax (Morel et al. 2020) (see more details below). The species tree was constructed by: aligning single-copy BUSCO genes with MAFFT v7.470 (--auto) (Katoh et al. 2002; Katoh and Standley 2013); trimming alignments with Trimal v1.4.rev22 (-automated1) (Capella-Gutiérrez et al. 2009); concatenating alignments into a super-matrix using FASconCAT v1.11 (Kück and Meusemann 2010); maximum-likelihood tree construction using IQTREE v2.0.6 (Hoang et al. 2018; Minh et al. 2020) after identifying the best-fitting phylogenetic model with the IQTREE2 Model Finder (Kalyaanamoorthy et al. 2017). The resulting tree was inspected to confirm that species and phyla relationships were compatible with the known literature and where necessary Mesquite v3.6.1 (Maddison and Maddison 2008) was used to correct branch positions. The species tree used in this chapter (available on GitHub) places sponges as sister-group to all other animals as this is one of the currently accepted scenarios (Feuda et al. 2017; McCarthy et al. 2023; Schultz et al. 2023). Furthermore, my previous work presented in Chapter 3 showed that no substantial difference was detected between sponge-first and ctenophore-first scenarios when performing gene-tree to species-tree reconciliations using a eukaryotic-wide set of organisms (see Supplementary Table S3.2).

Data mining

Enzymes for the retinol metabolism were chosen based on the pathway described on KEGG Database (KEGG map00830) (Kanehisa et al. 2021). Queries for BLASTP were collected from the KEGG Orthology lists (Kanehisa 2019) for each component of the pathway. BLASTP (Camacho et al. 2009) was conducted (with e-value threshold of 1e-5) for each query against the species database. To provide a preliminary annotation also for sequences from non-annotated non-model organisms, these were BLASTed versus the SwissProt Database (Poux et al. 2017) and the top hit was used as an approximate annotation.

Orthogroup inference

The results from BLASTP, organised by species, were used as "mini-proteomes" for orthogroup inference. By having reduced species proteomes by narrowing down to sequences with sequence similarity with the target enzymes of interest, it is in fact possible to reduce the computational load which is quite extensive for this type of analysis on large numbers of species. Two alternative methodologies for orthogroup inferences were used and compared in this work. In this way it was possible to verify the consistency of results when using different software. It also allowed to make sure not to miss out any potential sequences belonging to the orthogroups for the enzymes under investigation.

OrthoFinder

To insure best possible accuracy, OrthoFinder v.2.5.4 (Emms and Kelly 2015; Emms and Kelly 2019) was run with BLAST search (instead of default DIAMOND) and with the MSA workflow (using the default MAFFT for alignment and FastTree for tree inference). Furthermore, the species tree was provided (see above) rather than inferred by OrthoFinder. The inflation parameter used for MCL clustering was 1.3.

Broccoli

Broccoli v1.2.1 (Derelle et al. 2020) was run with kmer length for sequence clustering set to 80 to account for the distantly related species analysed; for the phylogeny step, maximum likelihood was chosen to maximise accuracy. Finally, regarding the species overlap parameter, several values were tested and finally the value of 0.9 was found to be

the best compromise between orthogroup accuracy (usually obtained with lower values) and avoidance of orthogroup fragmentation.

Filtering and annotation of orthogroups

To reach the goal of identifying orthogroups for the enzymes involved in the retinol metabolism, the orthogroups inferred by OrthoFinder and Broccoli must be annotated and potential unrelated orthogroups discarded. As a first step, all orthogroups that contained less than 4 sequences, or less than 4 species were discarded. Then, all sequences from each orthogroup were annotated using EggNog mapper (Cantalapiedra et al. 2021). One of the annotation fields outputted by EggNog is KEGG_pathways. Therefore, this was used to filter out any orthogroup that did not contain at least one sequence that obtained the KEGG map00830 (retinol metabolism) annotation. In this way it was possible to narrow down the number of orthogroups to analyse to identify orthogroups for our target enzymes. The remaining orthogroups were annotated by identifying the human sequences contained in them.

Comparison of OrthoFinder and Broccoli results and definition of final orthogroups

All enzymes known to be involved in the retinol metabolism were recovered as one or more orthogroup by both OrthoFinder and Broccoli. To assess the consistency between the results of the two methods, the next step was to compare the orthogroups by checking percentage of shared identical sequences amongst all OrthoFinder and Broccoli orthogroups (Figure 4.2). This comparison was visualised using Cytoscape v3.9.1 (Shannon et al. 2003), where orthogroups are represented as nodes and edges connecting the nodes represent the percentage of identical sequences shared between orthogroups. One-to-one correspondence with high percentage of identify was recovered in most cases and overall, it was possible to clearly establish the correspondence between OrthoFinder and Broccoli orthogroups. Final orthogroups used for subsequent phylogenetic analyses were the combined sequences collected with OrthoFinder and Broccoli. Cd-hit (Li et al. 2001; Fu et al. 2012) was used to remove duplicates with 100% identity after merging OrthoFinder and Broccoli orthogroups.

Reconstructing the evolutionary history for each orthogroup.

Phylogenetic Trees

A phylogenetic analysis was conducted for each orthogroup separately. Sequences from each orthogroup were aligned using MAFFT (--auto) (Katoh et al. 2002; Katoh and Standley 2013) and then trimmed using Trimal (with -gt 0.3 to remove columns with more than 70% gaps) (Capella-Gutiérrez et al. 2009). Resulting multiple sequence alignments were used for phylogenetic tree construction under maximum-likelihood using IQTREE2 (Hoang et al. 2018; Minh et al. 2020) after best-fit model testing (Kalyaanamoorthy et al. 2017).

Identifying clusters of orthologs with Possvm

The resulting gene trees were then further examined with Possvm (Grau-Bové and Sebé-Pedrós 2021), a tool that aids in identifying clusters of orthologs within gene trees facilitating the annotation process which, especially for large trees, can be very time consuming. A further advantage of this method is that it does not require a species tree as input for the ortholog sorting, eliminating potential biases related to disputed species relationships. Possvm was run using default parameters. As a result, each orthogroup corresponding to a broad enzyme family was further subdivided into smaller orthogroups corresponding to specific sub-families.

Reconstructing evolutionary events with GeneRax

Each gene tree was also reconciled to a species tree using GeneRax (Morel et al. 2020) enabling tree rooting and the discerning of speciation, duplication and loss events characterising each gene tree. The species tree used for reconciliation places sponges as sister-group to all other animals (see above) as this is one of the current accepted scenarios. Moreover, by comparing the reconciled trees with the Possvm-annotated tree, it is possible to control for potential inconsistencies and further investigate if the placement of sponges influenced them. Before running GeneRax, any polytomy in the gene trees were randomly resolved using ETE3 (Huerta-Cepas et al. 2016). GeneRax was run with the UndatedDL model that accounts for duplication and losses but not horizontal gene transfer events.

Data Availability

Additional supplementary material and raw output files are available at the GitHub repository: https://github.com/AAleotti/PhD_Thesis.

Acknowledgements

For this chapter I would like to thank Riccardo Kyriacou who during his time as a summer intern assisted me in optimising Broccoli parameters for orthogroup detection and then comparing Broccoli orthogroups with those from OrthoFinder using Cytoscape. My thanks also go to Julien Devilliers for his invaluable coding assistance, which facilitated the automation of various steps within this chapter.

References

Arne JM, Widjaja-Adhi MAK, Hughes T, Huynh KW, Silvaroli JA, Chelstowska S, Moiseenkova-Bell VY, Golczak M. 2017. Allosteric modulation of the substrate specificity of acyl-CoA wax alcohol acyltransferase 2. *Journal of Lipid Research* [Internet] 58:719–730. Available from: https://www.jlr.org/article/S0022-2275(20)33849-9/abstract

Arshavsky VY. 2002. Like Night and Day: Rods and Cones Have Different Pigment Regeneration Pathways. *Neuron* [Internet] 36:1–3. Available from: https://www.sciencedirect.com/science/article/pii/S0896627302009376

Batten ML, Imanishi Y, Maeda T, Tu DC, Moise AR, Bronson D, Possin D, Gelder RNV, Baehr W, Palczewski K. 2004. Lecithin-retinol Acyltransferase Is Essential for Accumulation of All-trans-Retinyl Esters in the Eye and in the Liver *. *Journal of Biological Chemistry* [Internet] 279:10422–10432. Available from: https://www.jbc.org/article/S0021-9258(17)47699-X/abstract

Bhatt-Wessel B, Jordan TW, Miller JH, Peng L. 2018. Role of DGAT enzymes in triacylglycerol metabolism. *Archives of Biochemistry and Biophysics* [Internet] 655:1–11. Available from: https://www.sciencedirect.com/science/article/pii/S0003986118301905

Blaner WS. 2017. Acyl-CoA wax alcohol acyltransferase 2: its regulation and actions in support of color vision1. *Journal of Lipid Research* [Internet] 58:633–635. Available from: https://www.jlr.org/article/S0022-2275(20)33841-4/abstract

Blaner WS, Das SR, Gouras P, Flood MT. 1987. Hydrolysis of 11-cis- and all-trans-retinyl palmitate by homogenates of human retinal epithelial cells. *Journal of Biological Chemistry* [Internet] 262:53–58. Available from: https://www.sciencedirect.com/science/article/pii/S0021925819758864

Blaner WS, Prystowsky JH, Smith JE, Goodman DS. 1984. Rat liver retinyl palmitate hydrolase activity. Relationship to cholesteryl oleate and triolein hydrolase activities. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* [Internet] 794:419–427. Available from: https://www.sciencedirect.com/science/article/pii/0005276084900080

Blomhoff R, Blomhoff HK. 2006. Overview of retinoid metabolism and function. *Journal of Neurobiology* [Internet] 66:606–630. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/neu.20242

Boussau B, Scornavacca C. 2020. Reconciling Gene trees with Species Trees. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.2:1-3.2:23. Available from: https://hal.science/hal-02535529

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* [Internet] 10:421. Available from: https://doi.org/10.1186/1471-2105-10-421

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Available from: https://www.biorxiv.org/content/10.1101/2021.06.03.446934v2

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* [Internet] 25:1972–1973. Available from: https://doi.org/10.1093/bioinformatics/btp348

Cheng JB, Russell DW. 2004. Mammalian Wax Biosynthesis: II. EXPRESSION CLONING OF WAX SYNTHASE cDNAs ENCODING A MEMBER OF THE ACYLTRANSFERASE ENZYME FAMILY *. *Journal of Biological Chemistry* [Internet] 279:37798–37807. Available from: https://www.jbc.org/article/S0021-9258(20)73052-8/abstract

Corbo JC. 2021. Vitamin A1/A2 chromophore exchange: Its role in spectral tuning and visual plasticity. *Developmental Biology* [Internet] 475:145–155. Available from: https://www.sciencedirect.com/science/article/pii/S0012160621000543

Derelle R, Philippe H, Colbourne JK. 2020. Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. *Molecular Biology and Evolution* [Internet] 37:3389–3396. Available from: https://doi.org/10.1093/molbev/msaa159

Dewett D, Labaf M, Lam-Kamath K, Zarringhalam K, Rister J. 2021. Vitamin A deficiency affects gene expression in the Drosophila melanogaster head. *G3 Genes|Genomes|Genetics* [Internet] 11:jkab297. Available from: https://doi.org/10.1093/g3journal/jkab297

Dewett D, Lam-Kamath K, Poupault C, Khurana H, Rister J. 2021. Mechanisms of vitamin A metabolism and deficiency in the mammalian and fly visual system. *Developmental Biology* [Internet] 476:68–78. Available from: https://www.sciencedirect.com/science/article/pii/S0012160621000762

Duester G. 2000. Families of retinoid dehydrogenases regulating vitamin A function. *European Journal of Biochemistry* [Internet] 267:4315–4324. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1432-1327.2000.01497.x

Edenberg HJ. 2007. The Genetics of Alcohol Metabolism: Role of Alcohol Dehydrogenase and Aldehyde Dehydrogenase Variants. *Alcohol Res Health* [Internet] 30:5–13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3860432/

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* [Internet] 16:157. Available from: https://doi.org/10.1186/s13059-015-0721-2

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* [Internet] 20:238. Available from: https://doi.org/10.1186/s13059-019-1832-y

Enright JM, Toomey MB, Sato S, Temple SE, Allen JR, Fujiwara R, Kramlinger VM, Nagy LD, Johnson KM, Xiao Y, et al. 2015. Cyp27c1 Red-Shifts the Spectral

Sensitivity of Photoreceptors by Converting Vitamin A1 into A2. *Current Biology* [Internet] 25:3048–3057. Available from: https://www.cell.com/current-biology/abstract/S0960-9822(15)01246-4

Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology* [Internet] 27:3864-3870.e4. Available from: https://www.sciencedirect.com/science/article/pii/S0960982217314537

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* [Internet] 28:3150–3152. Available from: https://doi.org/10.1093/bioinformatics/bts565

Grau-Bové X, Sebé-Pedrós A. 2021. Orthology Clusters from Gene Trees with Possvm. *Molecular Biology and Evolution* [Internet] 38:5204–5208. Available from: https://doi.org/10.1093/molbev/msab234

Hardie RC, Juusola M. 2015. Phototransduction in Drosophila. *Current Opinion in Neurobiology* [Internet] 34:37–45. Available from: https://www.sciencedirect.com/science/article/pii/S0959438815000173

Harrison EH. 2012. Mechanisms involved in the intestinal absorption of dietary vitamin A and provitamin A carotenoids. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* [Internet] 1821:70–77. Available from: https://www.sciencedirect.com/science/article/pii/S1388198111000849

Hirschberg J. 2001. Carotenoid biosynthesis in flowering plants. *Current Opinion in Plant Biology* [Internet] 4:210–218. Available from: https://www.sciencedirect.com/science/article/pii/S1369526600001631

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* [Internet] 35:518–522. Available from: https://doi.org/10.1093/molbev/msx281

Holmes RS. 2012. Vertebrate patatin-like phospholipase domain-containing protein 4 (PNPLA4) genes and proteins: a gene with a role in retinol metabolism. *3 Biotech* [Internet] 2:277–286. Available from: https://doi.org/10.1007/s13205-012-0063-7

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* [Internet] 33:1635–1638. Available from: https://doi.org/10.1093/molbev/msw046

Hussain Z, Uyama T, Tsuboi K, Ueda N. 2017. Mammalian enzymes responsible for the biosynthesis of N-acylethanolamines. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* [Internet] 1862:1546–1561. Available from: https://www.sciencedirect.com/science/article/pii/S1388198117301737

Jin M, Li S, Moghrabi WN, Sun H, Travis GH. 2005. Rpe65 Is the Retinoid Isomerase in Bovine Retinal Pigment Epithelium. *Cell* [Internet] 122:449–459. Available from: https://www.cell.com/cell/abstract/S0092-8674(05)00696-3

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* [Internet] 14:587–589. Available from: https://www.nature.com/articles/nmeth.4285

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Science* [Internet] 28:1947–1951. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3715

Kanehisa M, Sato Y, Kawashima M. 2021. KEGG mapping tools for uncovering hidden features in biological data. *Protein Science* [Internet] n/a. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4172

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* [Internet] 30:3059–3066. Available from: https://doi.org/10.1093/nar/gkf436

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* [Internet] 30:772–780. Available from: https://doi.org/10.1093/molbev/mst010

Kaylor JJ, Cook JD, Makshanoff J, Bischoff N, Yong J, Travis GH. 2014. Identification of the 11-cis-specific retinyl-ester synthase in retinal Müller cells as multifunctional O-acyltransferase (MFAT). *Proceedings of the National Academy of Sciences* [Internet] 111:7302–7307. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1319142111

Kienesberger PC, Oberer M, Lass A, Zechner R. 2009. Mammalian patatin domain containing proteins: a family with diverse lipolytic activities involved in multiple biological functions. *Journal of Lipid Research* [Internet] 50:S63–S68. Available from: https://www.sciencedirect.com/science/article/pii/S0022227520305885

Kramlinger VM, Nagy LD, Fujiwara R, Johnson KM, Phan TTN, Xiao Y, Enright JM, Toomey MB, Corbo JC, Guengerich FP. 2016. Human cytochrome P450 27C1 catalyzes 3,4-desaturation of retinoids. *FEBS Letters* [Internet] 590:1304–1312. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.12167

Kück P, Meusemann K. 2010. FASconCAT, Version 1.0, Zool. Forschungsmuseum A. Koenig, Germany, 2010.

Lamb TD. 2020. Evolution of the genes mediating phototransduction in rod and cone photoreceptors. *Progress in Retinal and Eye Research* [Internet] 76:100823. Available from: https://www.sciencedirect.com/science/article/pii/S1350946219301107

Lhor M, Salesse C. 2014. Retinol dehydrogenases: Membrane-bound enzymes for the visual function. *Biochem. Cell Biol.* [Internet] 92:510–523. Available from: https://cdnsciencepub.com/doi/abs/10.1139/bcb-2014-0082

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* [Internet] 13:2178–2189. Available from: https://genome.cshlp.org/content/13/9/2178

Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* [Internet] 17:282–283. Available from: https://doi.org/10.1093/bioinformatics/17.3.282

Maddison W, Maddison D. 2008. Mesquite: A modular system for evolutionary analysis. *Evolution* 62:1103–1118.

Mata NL, Radu RA, Clemmons RS, Travis GH. 2002. Isomerization and Oxidation of Vitamin A in Cone-Dominant Retinas: A Novel Pathway for Visual-Pigment Regeneration in Daylight. *Neuron* [Internet] 36:69–80. Available from: https://www.sciencedirect.com/science/article/pii/S0896627302009121

McCarthy CGP, Mulhair PO, Siu-Ting K, Creevey CJ, O'Connell MJ. 2023. Improving Orthologous Signal and Model Fit in Datasets Addressing the Root of the Animal Phylogeny. *Molecular Biology and Evolution* [Internet] 40:msac276. Available from: https://doi.org/10.1093/molbev/msac276

Meech R, Hu DG, McKinnon RA, Mubarokah SN, Haines AZ, Nair PC, Rowland A, Mackenzie PI. 2019. The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms. *Physiological Reviews* [Internet] 99:1153–1222. Available from: https://journals.physiology.org/doi/full/10.1152/physrev.00058.2017

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* [Internet] 37:1530–1534. Available from: https://doi.org/10.1093/molbev/msaa015

Moise AR, Kuksa V, Imanishi Y, Palczewski K. 2004. Identification of All-trans-Retinol:All-trans-13,14-dihydroretinol Saturase *. *Journal of Biological Chemistry* [Internet] 279:50230–50242. Available from: https://www.jbc.org/article/S0021-9258(20)67817-6/abstract

Moiseyev G, Chen Y, Takahashi Y, Wu BX, Ma J. 2005. RPE65 is the isomerohydrolase in the retinoid visual cycle. *Proceedings of the National Academy of Sciences* [Internet] 102:12413–12418. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0503460102

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution* [Internet] 37:2763–2774. Available from: https://doi.org/10.1093/molbev/msaa141

Nelson DR. 2018. Cytochrome P450 diversity in the tree of life. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* [Internet] 1866:141–154. Available from: https://www.sciencedirect.com/science/article/pii/S1570963917300857

Orland MD, Anwar K, Cromley D, Chu C-H, Chen L, Billheimer JT, Hussain MM, Cheng D. 2005. Acyl coenzyme A dependent retinol esterification by acyl coenzyme A:diacylglycerol acyltransferase 1. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* [Internet] 1737:76–82. Available from: https://www.sciencedirect.com/science/article/pii/S1388198105002210

Palczewski K, Kiser PD. 2020. Shedding new light on the generation of the visual chromophore. *Proc Natl Acad Sci U S A* [Internet] 117:19629–19638. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7443880/

Pingitore P, Romeo S. 2019. The role of PNPLA3 in health and disease. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* [Internet] 1864:900–906. Available from: https://www.sciencedirect.com/science/article/pii/S1388198118301458

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* [Internet] 33:3454–3460. Available from: https://doi.org/10.1093/bioinformatics/btx439

Redmond TM, Poliakov E, Yu S, Tsai J-Y, Lu Z, Gentleman S. 2005. Mutation of key residues of RPE65 abolishes its enzymatic role as isomerohydrolase in the visual cycle. *Proceedings of the National Academy of Sciences* [Internet] 102:13658–13663. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0504167102

Rowland A, Miners JO, Mackenzie PI. 2013. The UDP-glucuronosyltransferases: Their role in drug metabolism and detoxification. *The International Journal of Biochemistry & Cell Biology* [Internet] 45:1121–1132. Available from: https://www.sciencedirect.com/science/article/pii/S1357272513000654

Ruiz A, Winston A, Lim Y-H, Gilbert BA, Rando RR, Bok D. 1999. Molecular and Biochemical Characterization of Lecithin Retinol Acyltransferase *. *Journal of Biological Chemistry* [Internet] 274:3834–3841. Available from: https://www.jbc.org/article/S0021-9258(19)88015-8/abstract

Sahu B, Maeda A. 2016. Retinol Dehydrogenases Regulate Vitamin A Metabolism for Visual Function. *Nutrients* [Internet] 8:746. Available from: https://www.mdpi.com/2072-6643/8/11/746

Schreiber R, Taschler U, Preiss-Landl K, Wongsiriroj N, Zimmermann R, Lass A. 2012. Retinyl ester hydrolases and their roles in vitamin A homeostasis. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* [Internet] 1821:113–123. Available from: https://www.sciencedirect.com/science/article/pii/S1388198111000680

Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient gene linkages support ctenophores as sister to other animals. *Nature* [Internet]:1–8. Available from: https://www.nature.com/articles/s41586-023-05936-6

Seña C dela, Riedl KM, Narayanasamy S, Curley RW, Schwartz SJ, Harrison EH. 2014. The Human Enzyme That Converts Dietary Provitamin A Carotenoids to Vitamin A Is a Dioxygenase *. *Journal of Biological Chemistry* [Internet]

289:13661–13666. Available from: https://www.jbc.org/article/S0021-9258(20)38842-6/abstract

Seo M, Koiwai H, Akaba S, Komano T, Oritani T, Kamiya Y, Koshiba T. 2000. Abscisic aldehyde oxidase in leaves of Arabidopsis thaliana. *The Plant Journal* [Internet] 23:481–488. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-313x.2000.00812.x

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* [Internet] 13:2498–2504. Available from: https://genome.cshlp.org/content/13/11/2498

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* [Internet] 31:3210–3212. Available from: https://doi.org/10.1093/bioinformatics/btv351

Strauss O. 2005. The Retinal Pigment Epithelium in Visual Function. *Physiological Reviews* [Internet] 85:845–881. Available from: https://journals.physiology.org/doi/full/10.1152/physrev.00021.2004

Terao M, Romão MJ, Leimkühler S, Bolis M, Fratelli M, Coelho C, Santos-Silva T, Garattini E. 2016. Structure and function of mammalian aldehyde oxidases. *Arch Toxicol* [Internet] 90:753–780. Available from: https://doi.org/10.1007/s00204-016-1683-1

Thompson DA, Gal A. 2003. Vitamin A metabolism in the retinal pigment epithelium: genes, mutations, and diseases. *Progress in Retinal and Eye Research* [Internet] 22:683–703. Available from: https://www.sciencedirect.com/science/article/pii/S135094620300051X

Trifiletti RR. 2014. Vitamin A. In: Aminoff MJ, Daroff RB, editors. Encyclopedia of the Neurological Sciences (Second Edition). Oxford: Academic Press. p. 717–718. Available from: https://www.sciencedirect.com/science/article/pii/B9780123851574001172

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* 35:543–548.

Webb EC. 1992. Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* [Internet]. Available from: https://www.cabdirect.org/cabdirect/abstract/19930457289

Widjaja-Adhi MAK, Golczak M. 2020. The molecular aspects of absorption and metabolism of carotenoids and retinoids in vertebrates. *Biochim Biophys Acta Mol Cell Biol Lipids* 1865:158571.

Wong E, Anggono V, Williams SR, Degnan SM, Degnan BM. 2022. Phototransduction in a marine sponge provides insights into the origin of animal vision. *iScience*

[Internet] 25:104436. Available from: https://www.sciencedirect.com/science/article/pii/S2589004222007076

Zhao M, Ma J, Li M, Zhang Y, Jiang B, Zhao X, Huai C, Shen L, Zhang N, He L, et al. 2021. Cytochrome P450 Enzymes and Drug Metabolism in Humans. *International Journal of Molecular Sciences* [Internet] 22:12808. Available from: https://www.mdpi.com/1422-0067/22/23/12808

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).

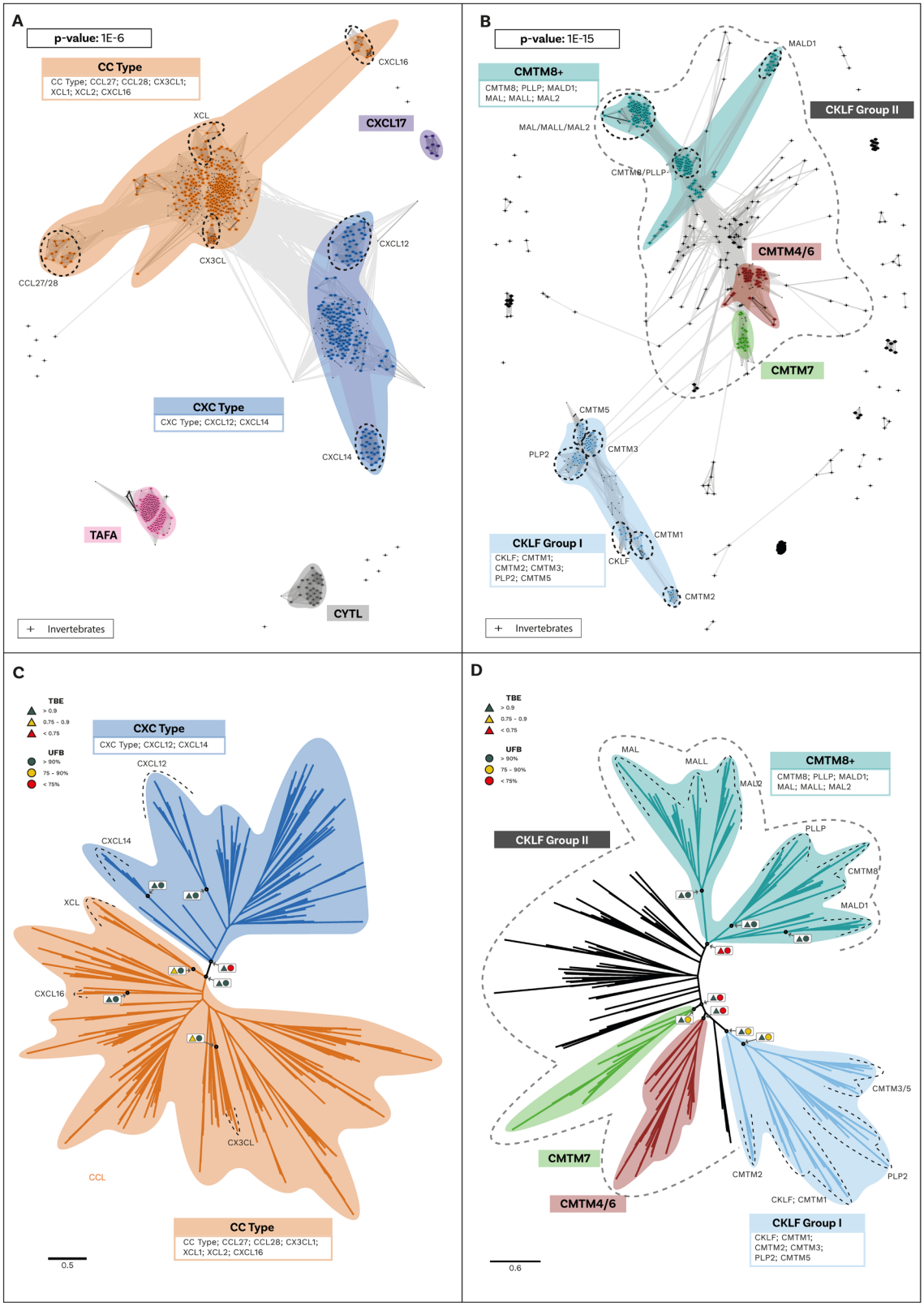


Figure 5.1. Cluster Analysis and Phylogeny of Ligand groups. (A) Similarity-based clustering, using CLANS, of canonical chemokines and related molecules with sequence similarity. Canonical chemokines

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.

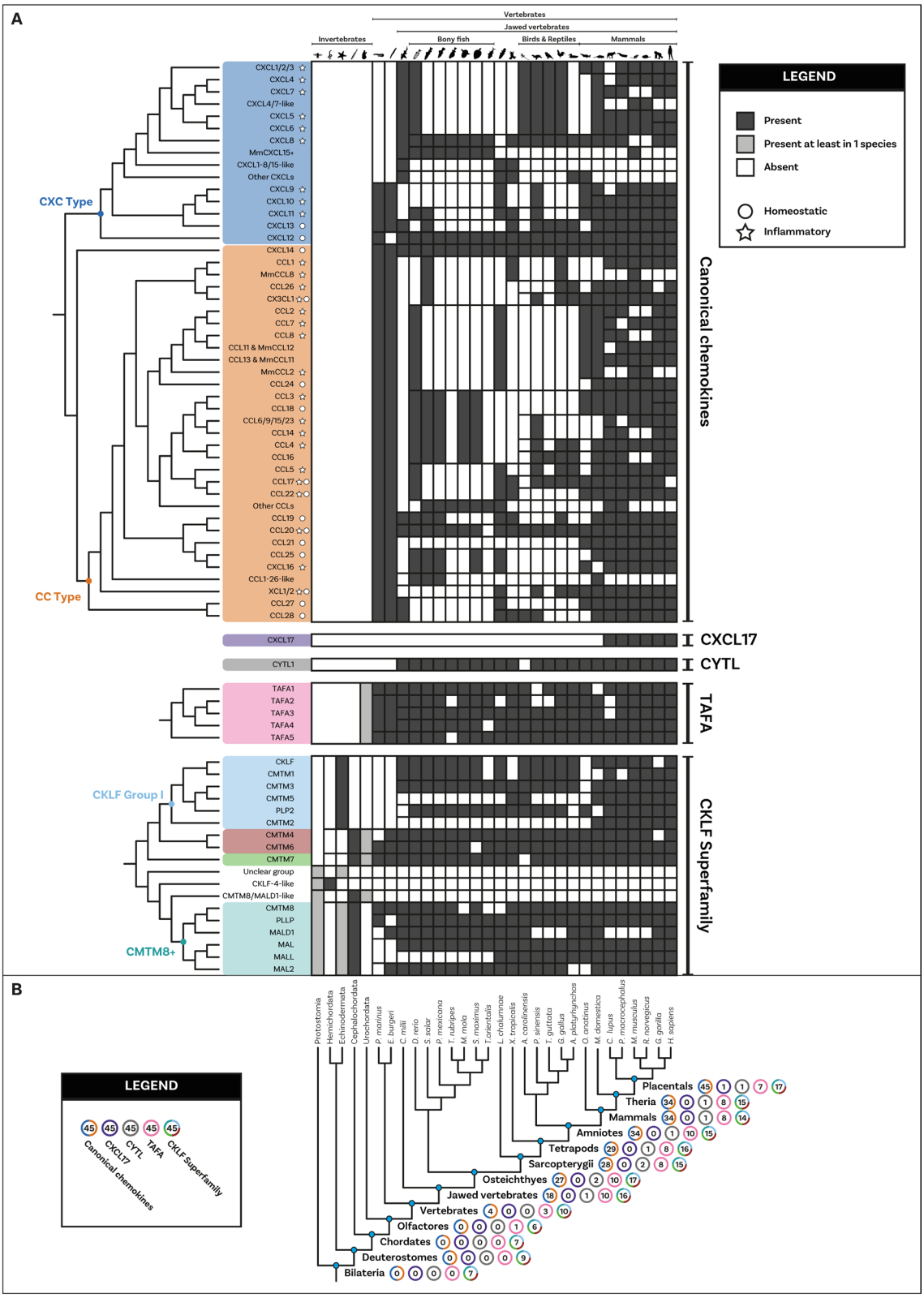


Figure 5.2. Distribution and duplication patterns of ligand groups. (A) Presence of all ligand groups are mapped onto a species tree. Gene trees and duplication events are based on the gene tree to species tree reconciliation analyses. The nomenclature for canonical chemokines is primarily based on known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

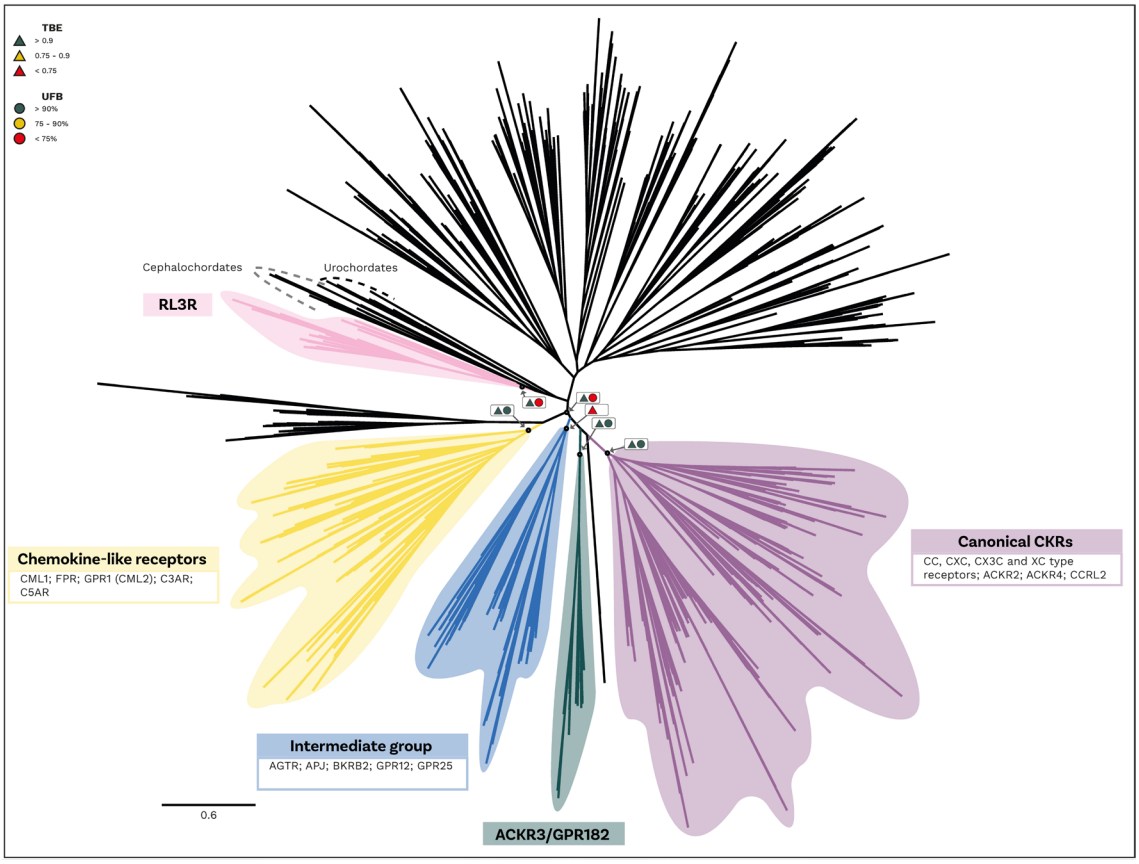


Figure 5.3. Phylogeny of Receptor groups. An unrooted maximum likelihood phylogeny of chemokine receptors. The tree shown is the transfer bootstrap expectation (TBE) tree including just the chordate specific clade from the ultrafast bootstrap tree (UFB). Node supports from both TBE (triangle) and UFB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

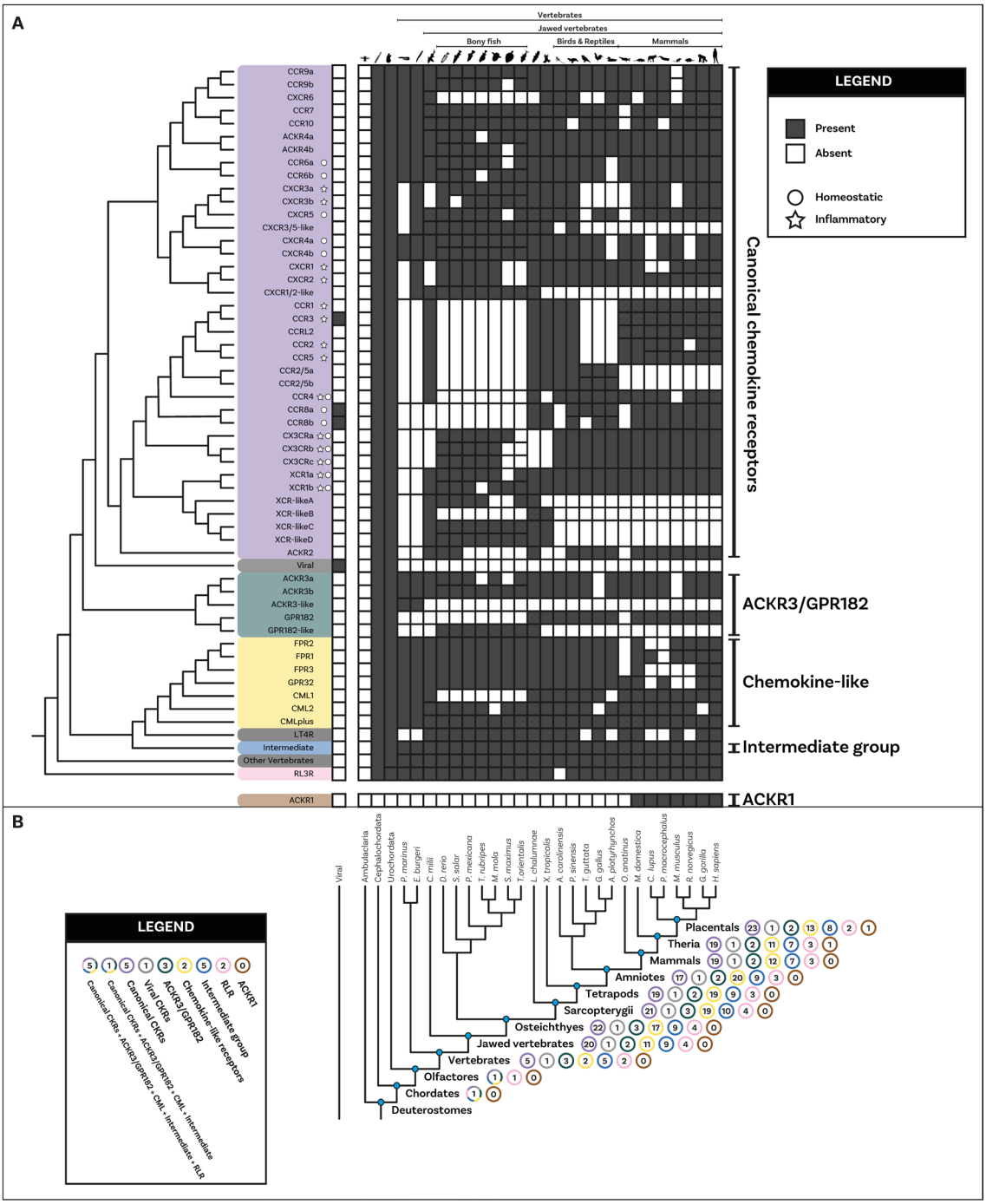


Figure 5.4. Distribution and duplication patterns of receptor groups. (A) Presence of all receptor groups are mapped onto a species tree. Gene trees and duplication events are based on the gene tree to species tree reconciliation analyses. The nomenclature for genes is primarily based on human chemokines. 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.

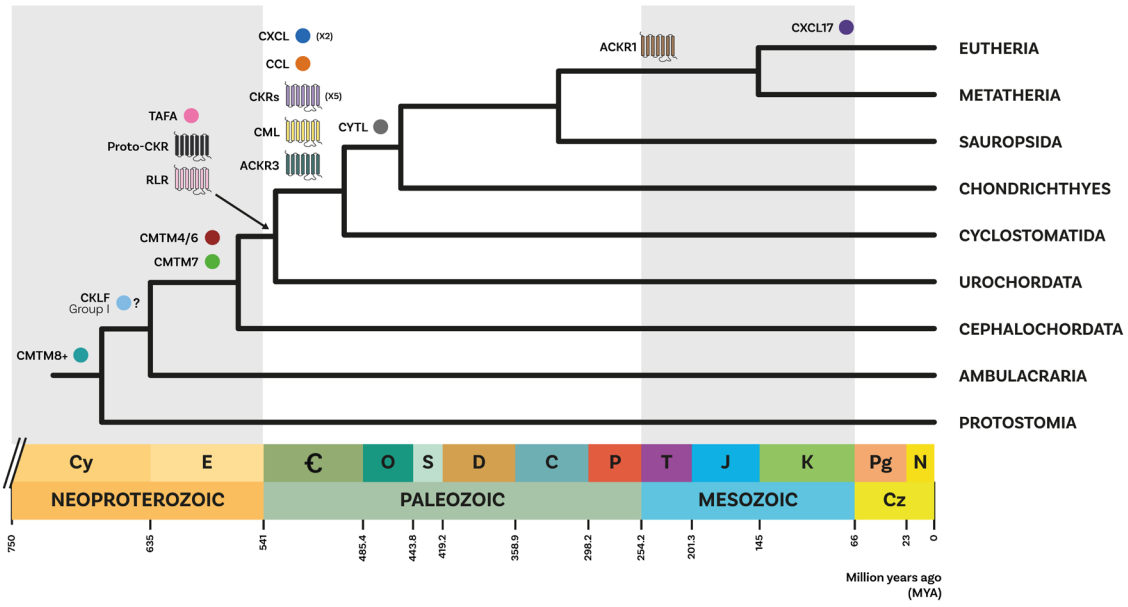


Figure 5.5. Summary of the evolution of ligands and receptors. A summary diagram of the evolution of the different chemokine system components. A simplified phylogenetic tree of species is shown, calibrated to time according to Dohrmann and Wörheide 2017 (72) for Deuterostomia and Bilateria nodes and 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).

Chapter 6

General Discussion, Conclusions and Future Perspectives

General Discussion

The power of evolutionary studies in understanding fundamental animal processes

Fundamental aspects of animal biology are deeply rooted in evolutionary history, especially the formative stages characterised by the major transition from a unicellular ancestor to a state of obligate multicellularity. The changes in genetic make-up that our early animal ancestors must have undergone are inextricably tied to this profound shift in lifestyle. The subdivision of specific tasks amongst different cells resulted in the origin of different cell types. Contemporarily, the need for cells of a multicellular organism to communicate and coordinate with each other both for internal physiology and for responding collectively to external stimuli, presented a critical challenge (Ruiz-Trillo and Nedelcu 2015). Such evolutionary pressures drove the expansion of gene families involved in cell signalling (Paps and Holland 2018). The evolution of vision, for instance, required a specialised cell type, the photoreceptor cell, capable both of detecting light, and of relaying this signal to other cells, that in turn can integrate this signal to elaborate a collective response. This was possible by coupling photosensitive molecules, the opsin bound to a cis-retinal molecule, with a phototransduction machinery capable of transducing the signal within the cell and culminating in ion channel modulation. The resulting flow of ions is responsible for the electrical signalling that starts the communication with other cells (Hardie and Juusola 2015; Lamb 2020). Similarly, cells of the immune system are responsible for maintaining the physiology of an organism both under normal conditions and when faced with external threats. Chemokine signalling, for example, plays an essential role in numerous processes in vertebrates by coordinating cell migrations throughout the body during development, homeostasis, and host defence. This system signals through small chemotactic proteins secreted by certain cells that are recognised by GPCR receptors on the target cells, which are induced to move along the gradient of these molecules (Murphy 2023).

Acknowledging this evolutionary history and recognising the centrality of cell signalling for the maintenance of multicellularity is essential for a deeper understanding of animal biology and evolution. This principle was an inspiration for the research presented in this

thesis, where I explored the evolution of vision and the chemokine system as examples of processes relying on cell signalling. To explore the evolutionary history of vision two primary aims were identified: first, investigating the evolution of photoreceptor cells by tracing the history of the phototransduction pathway components and the regulatory genes involved in the cell type specification; and second, reconstructing the evolution of the retinol metabolism pathway, which is responsible for the constant production of the photosensitive retinal that, when bound to opsin, enables the phototransduction process. To investigate the evolution of chemokine signalling, three main aims were pursued: first, clarifying the relationships between the "canonical" chemokines and chemokine receptors and "non-canonical" proteins that are also involved in the system; second, reconstructing the evolution of the ligand components; lastly, reconstructing the evolution of the receptor components. To address these aims, a combination of large-scale phylogenetic and bioinformatic methods were employed. Instead of limiting the focus to only a subset of molecular components, the analyses were conducted on a comprehensive set of components from the pathways and systems under investigation. This approach allowed to collect multiple pieces of evidence that combined created a broader picture of the trajectories of the systems under study, offering new insights into their evolution.

Animal-specific gene expansions as the molecular foundation of vision

The phylogenetic analyses of phototransduction gene families suggest an ancient origin in most cases. Many of these expansive families are present across Holozoa, with several of them being widely distributed throughout all eukaryotes. The involvement of these broad gene families in a variety of different pathways and processes other than vision, can explain their presence outside of animals. The few exceptions are opsins and G gamma, integral to both ciliary and rhabdomeric type phototransduction, along with the inhibitory subunit of the PDE6 enzyme and RGS9BP, specific to ciliary phototransduction. These families appear to be exclusive to animals, and in some cases to vertebrates. Within the extensive gene families, distinct subfamilies can be delineated. This granularity provides enhanced clarity as it enables a more detailed tracing of the evolutionary paths of those subfamilies containing genes known to be integral to phototransduction pathways in model organisms. A recurring observation, for instance, is that while the overarching gene family may span eukaryotes, numerous expansions

leading to diverse subfamilies predominantly took place within holozoans, just before the emergence of animals, or within the early history of animals. It is within these more recent subfamilies that we find the exact genes, that are well-documented in their roles in phototransduction of model organisms such as human and flies. While these observations generally apply to gene families of both major phototransduction pathways—ciliary and rhabdomeric—some differences emerged. For instance, within certain ciliary phototransduction families, the subfamilies most closely associated to vision are vertebrate-specific. This occurs for example in the super family of the PDE6 catalytic subunit, where the PDE6A/B/C is present only in vertebrates, and within the RGS super family, where RGS9 is present only in vertebrates. This, combined with the fact that the inhibitory subunit of PDE6 (PDE6G/H) appears to be vertebrate-specific implies that while foundational components of the ciliary pathway existed within all animals, some novelties, including expansions within older gene families and, possibly, *de novo* emergence of new genes, occurred concurrently with the evolution of vertebrates. While other members of the catalytic PDE6 subunit and RGS superfamilies likely can fulfil similar roles in non-vertebrates, the absence of PDE6G/H outside vertebrates suggests an auxiliary role. Indeed, in vertebrates this gene is involved in the shut-off step of phototransduction, contributing to improve the regulation and efficiency of the system in recovering from light stimuli (Lamb et al. 2018), rather than being essential for the basic phototransduction response.

In addition to offering valuable insights into when the full suite of phototransduction components coalesced in the genome of our ancestors, the exhaustive search for phototransduction genes across a diverse range of organisms enabled the compilation of specific markers for photoreceptor cells. These markers were subsequently used to detect photoreceptor cell profiles in the single-cell data of various animals, notably pinpointing potential photoreceptor candidates in all four non-bilaterian phyla. However, not all phototransduction families were consistently present in candidate photoreceptor cells of non-bilaterian species. Moreover, distinguishing between the prevalence of ciliary or rhabdomeric pathways proved challenging. This suggests that these phyla might not strictly adhere to the classical ciliary or rhabdomeric pathways. Instead, they could exhibit unique, lineage-specific variations. This, to some extent, seemed to be the case also for some bilaterian organisms such as sea squirt and sea urchin. Our current understanding of phototransduction is heavily influenced by a surprisingly narrow selection of

organisms. As such, it is plausible that these do not fully represent the true diversity of phototransduction systems across the animal kingdom. Research on the evolution of vision has, in recent years, seen an increase of interest in examining early-branching animals (Sebé-Pedrós, Saudemont, et al. 2018; Sebé-Pedrós, Chomsky, et al. 2018; Levy et al. 2021; Wong et al. 2022; McCulloch et al. 2023) —just like this study—as they offer key insights into the ancestral state of photoreceptor cells and vision. However, I argue that it is just as crucial not to overlook other bilaterian species. These could potentially serve as intermediaries, bridging our understanding between distantly related non-bilaterians and the handful of well-studied model organisms. It is only with a comprehensive grasp of the diversity of animal phototransduction pathways that it will be possible to fully trace the patterns that characterise their evolution. We are seeing exciting times, in which single-cell datasets of more and more animal species are being published (Gavriouchkina et al. 2022; Lust et al. 2022; Babonis et al. 2023; McCulloch et al. 2023; Piovani et al. 2023; Tominaga et al. 2023). Determining the evolutionary relationships among genes involved in phototransduction and pinpointing their presence or absence across a diverse spectrum of animals, is the first step for then exploring photoreceptor cell profiles in single cell data. This, combined with experimental studies, is contributing to further illuminating the intricate puzzle of vision evolution—a subject that has captivated researchers for centuries.

Similarly, a deeper understanding of the regulatory toolkits that characterise photoreceptor cells will also benefit from the availability of more and diverse single-cell datasets. However, interesting preliminary patterns are already apparent with the species that I was able to investigate in this thesis. Several families of regulatory genes frequently appeared in photoreceptor cells across animal species. These included, but were not limited to, transcription factors recognized for specifying the photoreceptor cell type in model organisms, like Six3/6, Meis2, and Tbx2 in humans, and the ATF4 ortholog crc and glass in flies. While, the exact combinations of these regulatory genes were not conserved throughout animals, certain families of transcription factors were more frequent than others, such as bZIP transcription factor, zinc finger C2H2 and homeobox families. This observation challenges the approach classically used of focusing on orthologous transcription factors as markers of shared photoreceptor cell profiles (Arendt 2003). Instead, a broader lens that focuses on transcription factor families might be more appropriate, especially given the challenges of identifying orthologs in distantly related

species. This alternative approach will benefit from large-scale phylogenetic analyses as well. Beyond transcription factors, which represent the most abundant type of regulatory genes shared across animal photoreceptor cells, other regulatory genes, including transcription cofactors and genes involved in chromatin conformation, were also consistently observed.

Finally, the last pieces of the puzzle that this thesis has to offer for understanding the evolution of vision, come from the phylogenetic analysis of the retinol metabolism enzymes that ensure the constant availability of cis-retinal, which bound to the opsin, is the first respondent to light stimuli. With this analysis, the retinol metabolism enzymes were found to belong to 12 distinct orthogroups—phylogenetically defined gene families. This confirmed some established family relationships but also revealed some surprises, such as the separation of the Diacylglycerol O-Acyltransferase (DGAT) enzymes into two distinct orthogroups. Overall, all orthogroups were revealed to have ancient origin and widespread distribution across eukaryotes. The only exception was the Lecithin Retinol Acyltransferase (LRAT) orthogroup, which exhibited a sporadic distribution both within and outside animals. Further refined phylogenetic analyses for each orthogroup yielded deeper insights into the substructures of these broad gene families as well as their distribution. An intriguing observation was that enzyme families with a marginal role in the pathway often had subgroups distributed widely across eukaryotes. Conversely, the subfamilies of the most specific enzyme families, such as RPE65, which catalyses the hydrolysis of stored all-trans-retinyl esters to 11-cis retinol, were generally exclusive to animals. Ultimately, the large-scale phylogenetic approach employed to reconstruct the evolution of both the phototransduction pathways, and the retinol metabolism unveiled a common narrative. It suggests that broad gene families, which might have originally played roles in foundational biological processes, experienced lineage-specific expansions within animals. Some of the emergent subfamilies were then co-opted into roles vital for vision, which, without them, would not have achieved its present-day complexity and sophistication.

Evolutionary dynamics of chemokine signalling

A similar approach to that used to investigate vision, was also applied to investigate the evolution of the chemokine signalling system. Though the system fundamentally consists

of two categories of components—ligands and receptors—it is complicated by the inclusion of "non-canonical" components. Additionally, chemokines, and to some extent their receptors, are small fast evolving molecules that have promiscuous interactions, thereby confusing their evolution. Employing sequence clustering techniques, my colleagues and I started by untangling the intricate relationships amongst canonical and non-canonical components. It was clarified that the non-canonical ligand families were not only distinct from canonical ones but also unrelated amongst themselves. In contrast, only one receptor group, the atypical chemokine receptor 1, was unrelated to all other receptors. These distinctions have important implications, as the literature has at times grouped disparate families based on their analogous functions and superficial sequence similarities (Tom Tang et al. 2004; Pisabarro et al. 2006; Weinstein et al. 2006; Tomczak and Pisabarro 2011; Chen et al. 2018; Liu et al. 2018; Zhang et al. 2018). While such categorizations may hold relevance in certain immunological contexts, they blur the true evolutionary relationships of these families. Furthermore, accurately discerning the relationships among these protein families is particularly important given their general pharmacological relevance and their key role as therapeutic targets (Lai and Mueller 2021).

Once the distinct groups of ligands and receptors were clarified, it was possible to examine their distribution across animals and investigate the details of each of their evolutionary histories. The strength of our research was the comprehensive examination of this system across a broad spectrum of animal species. While all canonical components were confirmed to be vertebrate-specific, several non-canonical components were found to be more ancient. For instance, TAFA ligands were detected also in urochordates, the sister group of vertebrates, and the CKLF super family was found across bilateria. This opens interesting avenues of future research in investigating the physiological roles of these molecules in invertebrates in the broader context of the evolution of immune systems. Finally, our detailed description of the pattern of duplication and loss events occurring within each gene family, provided an explanation for the diversity of the system. This ranged from uncovering ancient events, like the initial duplication at the stem of vertebrates which gave rise to all canonical and non-canonical receptors, to more recent occurrences such as the numerous mammalian-specific expansions observed in both chemokine and chemokine receptor groups.

Conclusions and Future Perspectives

By employing large-scale bioinformatic methods to investigate the origins and evolution of vision and chemokine signalling—two animal processes rooted in cell signalling— this thesis offered a multi-faceted perspective on the research questions. This approach led to various discoveries that can be further explored both bioinformatically and experimentally, laying foundation for future research in the field. Furthermore, this thesis provided an opportunity to experiment with various methodological strategies and bioinformatic tools tailored for expansive research questions, allowing for a deeper understanding of the advantages and drawbacks of each method. For instance, the curated strategy employed in Chapter 3 to identify the optimal protein profile for each individual gene family yielded highly accurate results. It also granted greater control over selecting gene families to analyse and offered flexibility in the scope of analyses, such as choosing between all GPCRs or focusing solely on opsins. Yet, this approach is also time-consuming, especially when investigating many and diverse gene families. Additionally, it necessitated a profound preliminary understanding of the targeted gene families to ensure informed decisions. On the other hand, utilizing algorithms that identify orthogroups, such those used in Chapter 4, is both rapid and robust. These tools are suitable for handling vast datasets. However, they offer less precision in determining the extent of the family, potentially including genes that might be too divergent from the original gene family of interest. The similarity-based clustering with CLANS used in Chapter 5 was in some respects an intermediary approach. While it was not as refined as the strategy of pinpointing family-specific profiles, and not as statistically robust as using orthogroup inferring tools, its adaptability in adjusting thresholds and intuitive visualization granted more control over determining the scope of the family collected. Ultimately, irrespective of the specific methods employed, for all gene families or orthogroups I conduct rigorous phylogenetic analyses, allowing me to detect any incongruences and place them in context. The lesson I learnt is that there is no single 'correct' method; it is instead essential is to consider the adequacy of the technique to the research question and the quantity and type of the gene families under study. To conclude, my comprehensive bioinformatic research on the evolution of the molecular components fundamental for vision and the chemokine system has provided a valuable starting point

to direct future bioinformatic research and to select targeted questions to explore experimentally.

References

Arendt D. 2003. Evolution of eyes and photoreceptor cell types. *Int J Dev Biol* 47:563–571.

Babonis LS, Enjolras C, Reft AJ, Foster BM, Hugosson F, Ryan JF, Daly M, Martindale MQ. 2023. Single-cell atavism reveals an ancient mechanism of cell type diversification in a sea anemone. *Nat Commun* [Internet] 14:885. Available from: https://www.nature.com/articles/s41467-023-36615-9

Chen K, Bao Z, Tang P, Gong W, Yoshimura T, Wang JM. 2018. Chemokines in homeostasis and diseases. *Cell Mol Immunol* [Internet] 15:324–334. Available from: https://www.nature.com/articles/cmi2017134

Gavriouchkina D, Tan Y, Ziadi-Künzli F, Hasegawa Y, Piovani L, Zhang L, Sugimoto C, Luscombe N, Marlétaz F, Rokhsar DS. 2022. A single-cell atlas of bobtail squid visual and nervous system highlights molecular principles of convergent evolution. *bioRxiv* [Internet]:2022.05.26.490366. Available from: http://biorxiv.org/content/early/2022/05/28/2022.05.26.490366.abstract

Hardie RC, Juusola M. 2015. Phototransduction in Drosophila. *Current Opinion in Neurobiology* [Internet] 34:37–45. Available from: https://www.sciencedirect.com/science/article/pii/S0959438815000173

Lai WY, Mueller A. 2021. Latest update on chemokine receptors as therapeutic targets. *Biochemical Society Transactions* [Internet] 49:1385–1395. Available from: https://doi.org/10.1042/BST20201114

Lamb TD. 2020. Evolution of the genes mediating phototransduction in rod and cone photoreceptors. *Progress in Retinal and Eye Research* [Internet] 76:100823. Available from: https://www.sciencedirect.com/science/article/pii/S1350946219301107

Lamb TD, Patel HR, Chuah A, Hunt DM. 2018. Evolution of the shut-off steps of vertebrate phototransduction. *Open Biology* [Internet] 8:170232. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rsob.170232

Levy S, Elek A, Grau-Bové X, Menéndez-Bravo S, Iglesias M, Tanay A, Mass T, Sebé-Pedrós A. 2021. A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity. *Cell* [Internet] 184:2973-2987.e18. Available from: https://www.sciencedirect.com/science/article/pii/S0092867421004402

Liu D-D, Song X-Y, Yang P-F, Ai Q-D, Wang Y-Y, Feng X-Y, He X, Chen N-H. 2018. Progress in pharmacological research of chemokine like factor 1 (CKLF1). *Cytokine* [Internet] 102:41–50. Available from: http://www.sciencedirect.com/science/article/pii/S1043466617303733

Lust K, Maynard A, Gomes T, Fleck JS, Camp JG, Tanaka EM, Treutlein B. 2022. Single-cell analyses of axolotl telencephalon organization, neurogenesis, and regeneration. *Science* [Internet] 377:eabp9262. Available from: https://www.science.org/doi/10.1126/science.abp9262

McCulloch KJ, Babonis LS, Liu A, Daly CM, Martindale MQ, Koenig KM. 2023. Nematostella vectensis exemplifies the exceptional expansion and diversity of opsins in the eyeless Hexacorallia. *EvoDevo* [Internet] 14:14. Available from: https://doi.org/10.1186/s13227-023-00218-8

Murphy PM. 2023. 15 - Chemokines and Chemokine Receptors. In: Rich RR, Fleisher TA, Schroeder HW, Weyand CM, Corry DB, Puck JM, editors. Clinical Immunology (Sixth Edition). New Delhi: Elsevier. p. 215–227. Available from: https://www.sciencedirect.com/science/article/pii/B9780702081651000150

Paps J, Holland PWH. 2018. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat Commun* [Internet] 9:1730. Available from: https://www.nature.com/articles/s41467-018-04136-5

Piovani L, Leite DJ, Guerra LAY, Simpson F, Musser JM, Salvador-Martínez I, Marlétaz F, Jékely G, Telford MJ. 2023. Single-cell atlases of two lophotrochozoan larvae highlight their complex evolutionary histories. :2023.01.04.522730. Available from: https://www.biorxiv.org/content/10.1101/2023.01.04.522730v2

Pisabarro MT, Leung B, Kwong M, Corpuz R, Frantz GD, Chiang N, Vandlen R, Diehl LJ, Skelton N, Kim HS, et al. 2006. Cutting Edge: Novel Human Dendritic Cell- and Monocyte-Attracting Chemokine-Like Protein Identified by Fold Recognition Methods. *The Journal of Immunology* [Internet] 176:2069–2073. Available from: https://www.jimmunol.org/content/176/4/2069

Ruiz-Trillo I, Nedelcu AM. 2015. Evolutionary Transitions to Multicellular Life: Principles and Mechanisms edited by Iñaki Ruiz-Trillo and Aurora M. Nedelcu. *Advances in Marine Genomics 2. Springer* [Internet] 91:370–371. Available from: https://www.journals.uchicago.edu/doi/abs/10.1086/688137

Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, Amit I, Hejnol A, Degnan BM, Tanay A. 2018. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat Ecol Evol* [Internet] 2:1176–1188. Available from: https://www.nature.com/articles/s41559-018-0575-6

Sebé-Pedrós A, Saudemont B, Chomsky E, Plessier F, Mailhé M-P, Renno J, Loe-Mie Y, Lifshitz A, Mukamel Z, Schmutz S, et al. 2018. Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* [Internet] 173:1520-1534.e20. Available from: https://www.sciencedirect.com/science/article/pii/S0092867418305968

Tom Tang Y, Emtage P, Funk WD, Hu T, Arterburn M, Park EEJ, Rupp F. 2004. TAFA: a novel secreted family with conserved cysteine residues and restricted expression in the brain. *Genomics* [Internet] 83:727–734. Available from: https://www.sciencedirect.com/science/article/pii/S088875430300332X

Tomczak A, Pisabarro MT. 2011. Identification of CCR2-binding features in Cytl1 by a CCL2-like chemokine model. *Proteins: Structure, Function, and Bioinformatics* [Internet] 79:1277–1292. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22963

Tominaga H, Nishitsuji K, Satoh N. 2023. A single-cell RNA-seq analysis of early larval cell-types of the starfish, Patiria pectinifera: Insights into evolution of the

chordate body plan. *Developmental Biology* [Internet] 496:52–62. Available from: https://www.sciencedirect.com/science/article/pii/S0012160623000179

Weinstein EJ, Head R, Griggs DW, Sun D, Evans RJ, Swearingen ML, Westlin MM, Mazzarella R. 2006. VCC-1, a novel chemokine, promotes tumor growth. *Biochemical and Biophysical Research Communications* [Internet] 350:74–81. Available from: https://www.sciencedirect.com/science/article/pii/S0006291X06020122

Wong E, Anggono V, Williams SR, Degnan SM, Degnan BM. 2022. Phototransduction in a marine sponge provides insights into the origin of animal vision. *iScience* [Internet] 25:104436. Available from: https://www.sciencedirect.com/science/article/pii/S2589004222007076

Zhang K, Shi S, Han W. 2018. Research progress in cytokines with chemokine-like function. *Cellular & Molecular Immunology* [Internet] 15:660–662. Available from: https://doi.org/10.1038/cmi.2017.121