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

Try to summarize key findings for the three main chapters..

### 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 phylogenetics and bioinformatics; and Flaviano Giorgini for his support and for providing useful inputs both scientifically and professionally.

I am extremely grateful to have been accompanied throughout these four years by awesome fellow PhD students. ‘’lunch breaks, parties, international dinners, games nights etc

I would like to thank particularly my little “siblings” Clifton, Matt, Julien, DaeNia and Frannie....

I would like to acknowledge my family and close friends for their encouragement throughout the years. Particularly my parents that have always gone above and beyond to support me.

### Table of Contents

[Abstract i](#_Toc149597339)

[Acknowledgements ii](#_Toc149597340)

[Table of Contents iii](#_Toc149597341)

[List of Figures v](#_Toc149597342)

[List of Tables vi](#_Toc149597343)

[Chapter 1 1](#_Toc149597344)

[General Introduction 1](#_Toc149597345)

[General Introduction 2](#_Toc149597346)

[The origin of multicellularity: a major evolutionary transition 2](#_Toc149597347)

[Expansion of signal transduction systems in animals 3](#_Toc149597348)

[General Aims of the Thesis 6](#_Toc149597349)

[The origin and evolution of vision in animals 6](#_Toc149597350)

[Aim 1: Reconstructing the evolution of the molecular components of photoreceptor cells. 8](#_Toc149597351)

[Aim 2: Reconstructing the evolution of the retinol metabolism. 8](#_Toc149597352)

[Evolution and molecular diversity of chemokine signalling systems. 8](#_Toc149597353)

[Aim 1: Uncovering the relationships among canonical and non-canonical components. 9](#_Toc149597354)

[Aim 2: Reconstructing the evolution of all canonical and non-canonical ligands. 9](#_Toc149597355)

[Aim 3: Reconstructing the evolution of all canonical and non-canonical receptors. 10](#_Toc149597356)

[References 11](#_Toc149597357)

[Chapter 2 18](#_Toc149597358)

[General Methods 18](#_Toc149597359)

[General Methods 19](#_Toc149597360)

[Phylogenetic analyses 19](#_Toc149597361)

[Dataset preparation 19](#_Toc149597362)

[Obtaining starting queries 19](#_Toc149597363)

[Choice of species 20](#_Toc149597364)

[Phylogenetic analyses 21](#_Toc149597365)

[Initial sequence similarity-based data mining 21](#_Toc149597366)

[Optimisation of final gene family datasets 22](#_Toc149597367)

[Annotating Sequences 22](#_Toc149597368)

[Multiple sequence alignment and trimming 23](#_Toc149597369)

[Inferring phylogenetic trees for each gene family 23](#_Toc149597370)

[Species trees 24](#_Toc149597371)

[Gene tree to species tree reconciliation 25](#_Toc149597372)

[Analyses of single-cell sequencing data 25](#_Toc149597373)

[Preliminary steps 26](#_Toc149597374)

[Choice of species and obtaining datasets 26](#_Toc149597375)

[Clustering cells into “metacells” 26](#_Toc149597376)

[Identifying photoreceptor cells and cross species comparisons 27](#_Toc149597377)

[Identification of candidates PRCs 27](#_Toc149597378)

[Exploration of the regulatory genes expressed in candidate PRCs 27](#_Toc149597379)

[Comparisons across species 27](#_Toc149597380)

[References 28](#_Toc149597381)

### List of Figures

### List of Tables

# 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