Chapter 4

The evolution of the retinol metabolism and implications for the origin of vision

Abstract

Introduction

The retinol metabolism comprises a series of enzymatic reactions that convert dietary vitamin A (retinol) into various bioactive compounds, primarily retinal for vision (REF) and retinoic acid for gene regulation (REF), ensuring the proper functioning of visual processes and other physiological roles in the body (REFS).

Retinol (Vitamin A1) 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 (REFS). Once in the cell, retinol is esterified to retinyl ester by the enzyme lecithin retinol acyltransferase (LRAT) (REF). When needed, retinyl ester is hydrolysed back to retinol (REF). Retinol is oxidized to retinal by retinol dehydrogenases (RDHs). Several other enzymes are involved in various steps of the retinol metabolism pathway as schematically shown in Figure 4.1 that summarizes 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 these enzymes 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 broad spectrum enzymes.

Retinal, particularly 11-cis-retinal, plays a crucial role in vision (REFS). 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 (REFS) (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 (REFS). 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 dehydrogenases (RALDHs). Retinoic acid serves as a signalling molecule that regulates gene expression and is critical for numerous developmental processes (REFS).

Retinol metabolism has been described mainly in human and ?. While little is known about its presence and composition in other organisms ?..

While most efforts to understand the details of all reactions and mechanisms comprising the retinol metabolism are mainly focused on mammalian/vertebrate models with the objective of improving human health/understanding of disease, and some degree of knowledge is available for other individual organisms like drosophila (?) plants (?), a comprehensive understanding of this pathway in the broader scale of all eukaryotic forms is missing.

Given the importance of the retinol metabolism, it is compelling to delve into its evolutionary history, especially when considering the broader 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.

Results and Discussion

Enzymes involved in retinol metabolism belong to 12 major orthogroups.

To understand the evolution of the retinol metabolism, I decided to reconstruct the evolution of all the enzymes involved in the pathway, as described by KEGG (Kanehisa et al. 2021) (Figure 4.1 and Table 4.1). To do this, I explored the genes encoding these enzymes in 101 species spanning all of Eukarya (Table 4.2 and its supp version with more info).

The first step to study the evolution of these genes was to first determine to which gene families or orthogroups they belonged to. In fact, while the number of enzymes participating in the pathway is relatively high, some of them might belong to a broader gene family, or to put it more precisely, orthogroup, i.e., a group of orthologs and paralogs deriving from the same original gene duplication.

Therefore, the first part of the analyses aimed to identify the orthogroups that the enzymes belong to. For this, a preliminary BLASTP (REF) was performed using sequences from the KEGG Orthology lists (Kanehisa 2019) as queries versus our database of 101 eukaryotes. A mention of the fact that the RPH component doesn’t have any seqs on kegg and therefore was discarded from further search until future work can provide at least some preliminary seqs to work with. The results of this blast were used as input for orthogroup identification pipeline. The details can be found in the Methods section, but briefly, two alternative software (Broccoli (REF) and Orthofinder (REF)) were used to independently assess orthogroups, then the results were compared, and consensus groups were defined. Briefly mention key differences of methods?

The results of the orthogroup identifications and the comparison between the two methods is shown in Figure 4.2. First of all, we can see how the two methods are largely consistent, with many cases of one-to-one correspondence of orthogroups. However, it is also immediately noticeable that Orthofinder tended to provide fewer and larger orthogroups, while Broccoli provided more and in some cases smaller orthogroups. As a consequence, some gene families appeared fragmented into multiple smaller orthogroups according to Broccoli only. Quickly mention that we found a GK group (or subgroup) but this gene has nothing to do with retinol metabolism so excluded from further research. Table 4.3 summarizes the final 12 orthogroups identified and shows the comparison of Orthofinder and Broccoli with each other and the original KEGG groups.

Overall we identified some interesting and unexpected findings: such as that DGAT and DGAT2L4 are not to be considered the same gene family/orthogroup (according to both Orthofinder and Broccoli) and that the SDR and RDH families are intermingled, possibly indicating that they belong to a broader orthogroup. For the latter, to discriminate more rigorously the relationship between SDR and RDH, all orthougroups were collected as one big orthogroup for phylogenetic analysis.

Possvm and Generax describe substructure of orthogroups.

Aim of this section.

Vague mention of the phylogenetic trees themselves before possvm and generax.

Differences between two methods. What does possvm do, what does generx do. What are pros and cons of each. (Does possvm output a rooted or unrooted tree). Possvm is fast, easy to run (from laptop?) and easy to annotate.. generax is slow but very precise (?)

“As the species overlap algorithm relies on the implicit taxonomic information contained in the gene tree’s topology, this approach is suitable for cases where the species tree is unknown or unavailable.” – Possvm paper. 😊

The details for each OG, 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). This enzyme appears to be involved only in retinol metabolism (Table 4.1) according to the KEGG Database (REF), 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 consensus RETSAT orthogroup contained 338 sequences distributed throughout all major eukaryotic clades (Figure 4.3A).

Ortholog sorting with Possvm identified 7 orthogroups within the RETSAT family, with one orthogroup containing RETSAT together with the related enzyme PYRD2 as well as CTR enzymes (Figure 4.3B). Gene tree to species tree reconciliation with GeneRax revealed a high number of events (especially losses) in proportion to the size of the orthogroup (Figure 4.3C). More details about tree topologies?

Check out Weber et al 2020 for further discussion?

**PNPLA4**

The Patatin Like Phospholipase Domain Containing 4 (PNPLA4) enzyme plays a role in the hydrolysis of retinyl esters to retinol (Schreiber et al. 2012). 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 consensus orthogroup contains 215 sequences. While being present in both major eukaryotic clades, this orthogroup appears to be missing in basal groups of Amorphea, such as the Holomycota branch that includes Fungi (Figure 4.4A).

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 tight relationship between PNPLA4 and other PNPLA genes is in accordance with evidence suggesting that some of them are also involved in retinol metabolism (REF).

**ALDH1**

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

ALDH1 is identified as its own orthogroup by both OrthoFinder and Broccoli (Figure 4.2) and the consensus orthogroup consists of 765 sequences. It is a ubiquitous orthogroup with only a handful of eukaryotic species missing it (Figure 4.5A).

Possvm identified a complex substructure within the ALDH1 orthogroup, subdividing it into 44 orthogroups (Figure 4.5B). 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). More details?

**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). 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 also 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 consensus orthogroup consists in 322 sequences. We find this orthogroup to have a patchy presence throughout most eukaryotic clades (Figure 4.6A).

Possvm analyses of orthologs identified 16 orthogroups within this family, with BCO1, RPE65, as well as BCO2, belonging to the same orthogroup (Figure 4.6B). GeneRax recovers a consistent topology and a moderately high number of events (Figure 4.6C). More info?

**LRAT**

Lecithin Retinol Acyltransferase (LRAT), also known as Phosphatidylcholine--Retinol O-Acyltransferase, catalyses the esterification of all-trans-retinol into all-trans-retinyl ester (REF). It belongs to three KEGG pathways (Table 4.1).

OrthoFinder and GeneRax orthogroups for this enzyme correspond 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. GeneRax confirms the same tree topology and identifies a rather high number of events relative to the number of sequences (Figure 4.7C). More info?

**RDH/DHRS**

Retinol Dehydrogenase (RDH) enzymes are responsible for the oxidation of retinol to retinal (REF). RDH5 in particular is responsible for the conversion of 11-cis-retinol to 11-cis-retinal, the visual chromophore (REF). 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 (REF). Other enzymes that also belong to this family are Dehydrogenase/Reductases SDR family (DHRS), of which several are also implicated in the 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 approaches identified two main orthogroups, one that contains exclusively RDH enzymes and the other that includes a mixture of RDH and DHRS enzymes. However, on top of those, Broccoli identified numerous other small orthogroups some with a more RDH type profile and other more DHRS specific. Two of the Broccoli orthogroups even share a very small number of sequences with the GK orthogroup which was collected at the BLASTP step but, 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), meaning that we can safely consider ADH as a separate 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 substructure identified by Possvm. 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 in turn clading with other members of this large family. GeneRax recovered a largely compatible substructure and revealed a very large number of evolutionary events even when correcting 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, rather monophyletic clades within this broad gene family include enzymes that have been described (based primarily on structure and function) to belong to different subfamilies, indicating that a phylogenetic approach was needed to clarify the evolutionary relationship between these enzymes. Finally, as mentioned RDH and DHRS families have been described to belong to the vast SDR super family, therefore, it is possibile that further details of subfamily relationships could be revealed by broadening the analysis to other SDR members. However, that would present an extremely challenging task as already this orthogroup touched the maximum scale of sequences with which this type of detailed phylogenetic analysis is feasible.

(Figure is collapsed in such a way to keep one clade per possvm orthogroup.)

**DGAT1**

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

KEGG proposes that both DGAT1 and DGAT2L4 (see below) occupy the same position in the pathway (Figure 4.1 and Table 4.1); however, 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 Eukarya 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 amount of evolutionary events within this orthogroup (Figure 4.9C). More details?

**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 (Figure 4.1) (Kaylor et al. 2014; Arne et al. 2017; Blaner 2017). 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 topology is maintained in the reconciled tree by GeneRax that calculated quite a high number of events (Figure 4.10C). More details?

**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 (REF). The number of CYP enzymes is so vast that it is generally considered to be a super family in turn subdivided into families and subfamilies (Nelson 2018). For example, the CYP27C1 enzyme, the most specific to the retinol metabolism (Table 4.1), belongs to the family 27, subfamily C, and is the member 1. It catalyses the 3,4 desaturation of all-trans-retinol to all-trans-3,4-didehydroretinol (REF) (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 missing 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. More details?

**AOX**

Aldehyde Oxidase 1 (AOX1) is responsible for the oxidation of a wide variety of aldehydes to their corresponding carboxylic acids (REF). Within the retinol metabolism it is able to oxidise retinal to retinoate (REF) (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 and 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 (REF). The total orthogroup of AOX includes 599 sequences. It is overall present in all eukaryotes with exception of some clades, 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). More details?

**ADH**

Alcohol dehydrogenase (ADH) enzymes play crucial roles in the metabolism of alcohols (REFS). In the retinol metabolism ADHs play a role in the oxidation of retinol to retinal (REF). Although RDH is the primary enzyme for this reaction, especially within the retina, ADHs can also contribute and within humans are especially used in non-visual related tissues such as the liver (REFS). 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 that contained mainly ADH, and the other that was a mixture that included 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 number 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). More info?

**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 (REF). 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 (REFS). 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).

Our phylogenetic analysis uncovers that UGT1 and UGT2 are more related to each other and UGT3 and UGT8 are more related to each other, but that all of them belong to one monophyletic clade that 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 amount of events relative to number of sequences, all of which indicates that the UGT orthogroup is quite compact. More info?

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 (REF). 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 (REF). Thus, understanding the origins and evolution of vision necessitates exploring the evolution of retinol metabolism, which sustains our light sensitivity.

In this chapter, I addressed the issue of understanding the evolution of the retinol metabolism by first, understanding the relationship between the enzymes involved in the pathway and categorising them in broad gene families or orthogroups, second by understanding the distribution of these orthogroups throughout eukaryotes and lastly by describing the evolutionary events that characterise the history of each orthogroup.

Enzymes are traditionally grouped into families based primarily on their function (types of reactions catalysed) and to some degree tertiary structures (REFS). With my orthogroup analyses I aimed to investigate whether the enzymes involved in the retinol metabolism could be categorised based on evolutionary relationships and how this related to known enzymatic families. I found that the enzymes involved in the retinol metabolism belong to 12 distinct orthogroups (Figure 4.2, Table 4.3), with some being involved in some of the most crucial steps for the recycling of 11-cis-retinal and others playing more marginal roles (Figure 4.15A). The analysis largely recapitulated the known enzymatic families but revealed some interesting insights. For example, Diacylglycerol O-Acyltransferase enzymes have been traditionally grouped into one big family, here I found sufficient evidence to suggest that there is enough evolutionary distance between DGAT1 on the one hand and DGAT2 and DGAT2-like molecules such as DGAT2L4 to be considered separate orthogroups. Another example involves the enzymes involved in the reaction that transforms retinol to retinal that is done by some subfamilies of the broad SDR family. While current nomenclatures suggest a distinction between RDH and DHRS, the orthogroup analyses suggested a more complex set of relationships which was then later confirmed through phylogenetic analyses (Figure 4.8). Ultimately all this suggests that within the SDR family, phylogenetic relationships would define different subfamilies compared to the currently established ones.

Regarding the distribution of orthogroups throughout eukarya, the first consideration is that all orthogroups appear to be very ancient spanning most eukaryotic clades (Figure 4.15B). The only exception would be LRAT that appears to be present primarily in animals (except placozoans and ctenophores) and in a handful of other species. Several orthogroups including some very specific ones to the retinol metabolism, like RETSAT and ALDH1, are indeed present in all major clades, although only one orthogroup (RDH/DHRS) was present in every single species examined (Figure 4.8). Amongst animals specifically, only placozoans and ctenophores appear to miss some of the orthogroups including BCMO1/RPE65 and LRAT, that are some of the most specific enzymes of the pathway. This is interesting for two reasons; first this may indicate that these two phyla may have an alternative version of retinol metabolism in which these enzymes are substituted by other enzymes (e.g. LRAT’s job can also be done by DGAT1 and DGAT2L4); second the fact that conversely sponges, which lack opsins, do possess all orthogroups (although not all species all orthogroups) has interesting implications in understanding the origin of vision, reinforcing the growing interest in understanding potential light sensing processes in sponges with alternatives to opsins (chapter 3 and refs).

Regarding the description of evolutionary events within each orthogroup, the approach of using both Possvm and GeneRax allowed on the one hand a quick and easy way to defined sub-orthogroups as proxy of the fragmentation or not of the orthogroup and on the other hand a detailed and rigorous description of events of duplication, speciation and losses as a proxy of how much evolutionary events have happened as opposed to a linear history. Many events can also explain high diversity of subfamilies. These descriptions allowed to identify which orthogroups presented the most complex evolutionary histories such as ADH that had one of the highest number of orthogroups per sequences as well as one of the highest number of events per sequences. It also discriminated between large orthogroups with straightforward subgroups like CYP versus large orthogroups with complex substructure like RDH/DHRS. Any other interesting stats or examples?

To conclude, the ancient origin of the enzymes involved in the replenishment of 11-cis-retinal point towards a situation in which this molecular set up was already in place long before the advent of vision during the early evolution of animals. This opens a whole new set of questions regarding the details of the evolution of vision. For example, understanding if the enzymes bioinformatically identified in this study in early branching animals do indeed function in physiological setting to complete the retinol pathway and recycle 11-cis-retinal (e.g. even in sponges were in theory it shouldn’t be important since the don’t have opsins anyway). Furthermore, in humans several known cell types are involved in different steps of the pathway e.g., RPE for isomerization of all-trans-retinol to 11-cis-retinol and subsequently to 11-cis-retinal, and Mueller cells, for the uptake and release of retinol facilitating the movement between PRC and RPE. Therefore it would be interesting to understand if equivalent potentially homologous cell types occur in early branching animals and/or if parts of the pathway are carried out in unrelated cell types or within one cell type.

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 (REF) 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; 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 (REFS) 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 (REFS) 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 exploited 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 of corresponding 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 subfamilies.

**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: put link.

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* 58:719–730.

Blaner WS. 2017. Acyl-CoA wax alcohol acyltransferase 2: its regulation and actions in support of color vision1. *Journal of Lipid Research* 58:633–635.

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* 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* 25:1972–1973.

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* 279:37798–37807.

Duester G. 2000. Families of retinoid dehydrogenases regulating vitamin A function. *European Journal of Biochemistry* 267:4315–4324.

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* 27:3864-3870.e4.

Grau-Bové X, Sebé-Pedrós A. 2021. Orthology Clusters from Gene Trees with Possvm. *Molecular Biology and Evolution* 38:5204–5208.

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* 1821:70–77.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 35:518–522.

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

Jin M, Li S, Moghrabi WN, Sun H, Travis GH. 2005. Rpe65 Is the Retinoid Isomerase in Bovine Retinal Pigment Epithelium. *Cell* 122:449–459.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589.

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Science* 28:1947–1951.

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* 30:3059–3066.

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

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* 111:7302–7307.

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

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

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* 99:1153–1222.

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* 37:1530–1534.

Moise AR, Kuksa V, Imanishi Y, Palczewski K. 2004. Identification of All-trans-Retinol:All-trans-13,14-dihydroretinol Saturase \*. *Journal of Biological Chemistry* 279:50230–50242.

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* 102:12413–12418.

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* 37:2763–2774.

Nelson DR. 2018. Cytochrome P450 diversity in the tree of life. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1866:141–154.

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* 33:3454–3460.

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* 102:13658–13663.

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* 1821:113–123.

Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient gene linkages support ctenophores as sister to other animals. *Nature*:1–8.

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* 289:13661–13666.

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.* 13:2498–2504.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

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.