Methods – General

To address the two broad questions of this thesis – the evolution of vision and the evolution of chemokine signalling – I used bioinformatic approaches. Although the exact methods used for each chapters are specific, they share some general communalities.

In this chapter, I will cover the main steps that were common for all results chapters, describing how the techniques/tools used work. However, further details of how these techniques/tools were used are described more precisely in each respective chapters.

Two broad categories of methods were used in this thesis: molecular phylogenetic analyses; and single-cell RNA sequencing analyses.

**Molecular phylogenetic analyses (used in chapters 3,4,5)**

* A first paragraph mentioning the main steps with preliminary explanation:
  + Identifying what gene families to study to answer research question e.g. from kegg pathways etc.
    - Kegg is a tool I widely using as starting point for pathways/metabolisms
    - Put as figure the kegg map with all pathways.
  + Choice of species to look into
  + Building phylogenetic trees
  + Reconciling gene tree to the species tree
* Choice of species to do the search in:
  + Taxonomic distribution
  + BUSCO proteome completeness
  + Differences in different projects: chapters 3 and 4 broad eukaryotic group; chapter 5 animals.
* Phylogenetic analyses:
  + Data mining and Identification of gene families:
    - Often BLAST is enough: either blastp or psi-blast: explain vaguely the difference. (can cite phylogenetic handbook...). Considerations about the e-value threshold.
    - Where necessary further steps may be necessary before proceeding to phylogenetic tree reconstruction. These can be more targeted/specific or can rely on statistical software to detect best groups.
    - In this thesis methods used include:  
      - chapter 3: targeted sequence filtering based on known protein domains expected in protein family.

- chapter 4: use of orthogroup inferring software to detect software.

- chapter 5: cluster analysis of sequences to visualise clusters: simpler approach than orthogroup inference but good starting point for then doing phylogenetic analysis. A sort of in between strategy compared to previous two chapters.

* + Annotating sequences:
    - Necessary because not all species have annotated protein sequences. Annotation helps to navigate cluster of sequences and trees.
    - Most commonly used method is to do a blast vs swissprot and keep best hit. Good especially for model organisms like human and mouse, but performs less well for non-model organisms. Furthermore, where necessary while understanding the structure withing a tree, further manual annotation of some sequences that were key to annotate a specific clade was carried out. This would often take advantage of species-specific databases, such as genecards for human, xx for mouse, flybase for flies, echinobase for sea urchins and other echinoderms and yy for Arabidopsis thaliana.
  + Aliments and trimming: what are considerations to make. Pros and cons of different parameters (e.g. when you have short sequences).
  + Gene tree construction: model finder, iqtree, support values (mainly UFB, TBE, with others being bayes and lrht?)
    - Quick definition of maximum likelihood
  + Species tree – necessary for some subsequent steps in some cases.
    - Constructing species tree as backbone for gene tree: using busco genes for supermatrix; checking for known topology etc.
  + Gene tree to species tree reconciliation:
    - Preparing gene tree: no polytomies
    - Generax: parameters
    - Ezio comment: still not clear how reconciliation works.. in the thesis can be explained in the methods chapter.. in the paper does it need additional explanantion? 🡨 see explanation I inserted in chapter 3.

**Single-cell analyses (used in chapter 3)**

* Choice of datasets / considerations... although generic because details should be in the chapter.
  + Species choice
  + Different life stages of tissues 🡨 really a limitation rather than a choice.
* Choice of method to cluster cells:
  + First some background knowledge of options
  + Then talk about my choice: metacells because of low coverage...
* Issue of having to compare distantly related species
  + As well as not always same tissue. E.g. in some cases whole body in others the retina.
* Tailored methods / question driven methods