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 exact methods are described precisely in each respective chapter, several basic approaches were shared amongst the different projects. These methodologies provide the basis for my research and include genome assessment analyses, cluster-based analysis, phylogenies, and gene-tree to species-tree reconciliations. In this chapter I will provide an overview of the common methodologies, which will serve as a common foundation for the next chapters.

Molecular phylogenetic analyses

All aims of this thesis required phylogenetic analysis of gene families essential to he biological process of interest. Here are the major steps common to Chapters 3, 4, and 5. With a few differences.

**Preliminary steps**

***Obtaining starting queries***

It might sound obvious, but one of the preliminary steps is to identify which gene families to investigate with phylogenetic studies. A starting point is of course the literature, however, when studying large pathways, to not miss out on important components it can be sensible to rely on pathways databases as a starting checklist for all genes of interest. A reliable and comprehensive database of pathways is KEGG, that also provides lists of known homologs for components of the pathways. I often use this resource as a starting point for collecting sequences of references, or queries, to be used as a starting point for my subsequent data mining steps.

***Choice of species and species trees***

The comparative analysis of systems and signalling pathways in different organisms requires of the examination of genomes and predicted proteomes across diverse species. Therefore, another preliminary step is the choice of species to look into. This is a crucial step for various reasons. First, the correct taxonomic sampling depends on the research question. For example, when investigating fundamental pathways that likely utilise some ancient components, it is important to include distantly related species. For example, chapters 3 and 4 ranged all of eukayra. Conversely, when studying the evolution of a taxon-specific system, e.g., chemokine signalling, a more restricted set of organisms will do the job and it is just important to include some outgroups. E.g. ingroup is vertebrates, potentially bilaterians and non-bilaterians are outgroup. During the course of my phd studies I used as my primary references for eukaryote classification and phylogenetic placements Adl 2019 and Burki 2020. The second reason why this step is crucial is because we want to make sure that the genomes/proteomes used in our study are high quality. The quality of a genome, transcriptome or predicted proteome sequence 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 accurate representation of an organism’s genetic blueprint. This is critical for identifying and annotating genes correctly, mapping transcripts, and predicting protein sequences (Simakov et al. 2022). 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).

The primary characteristic of a high-quality proteomes is that they are complete. That way if we do not find a gene family we can be confident that it is because it is missing from the lineage and not due to technical issues. Ideally. In some cases there are some key species/taxa that we want to include in our study even if their proteome is not very complete. What we can do in these cases is to include multiple closely related species to maximise our chances of finding presence of gene families in that taxonomic lineage. The tool used throughout my phd studies to assess the completeness of proteomes was BUSCO. This tool searches for a list of genes that are well known (benchmarked) to be present in a given taxon in single copy. Therefore the BUSCO analysis can be tailored by taxon: if I am including all eukaryotes I will assess the completeness of genes that are supposed to be present in single copy in eukaryotes, if I am including only animals then I look for BUSCO genes characteristic only of animals. The BUSCO analyses gives a percentage of BUSCO genes that were found and in single copy. It also gives percentage of BUSCO genes found in multiple copies, this is also a useful parameter to keep into consideration as it may in some cases be an indicator of assembly issues of the genome. Finally it gives percentages of fragmented and missing BUSCOs. With these assessments it is possible to make a final species database combining both taxonomic and proteome completeness considerations.

To approach this, I made the analysis of my previously obtained database with the BUSCO (Benchmarking Universal Single-Copy Orthologs), this tool played an instrumental role in assessing the quality and completeness of the genomic, transcriptomic, and predicted proteome data that I employed. BUSCO is known for its ability to evaluate the integrity of these datasets, facilitating the identification and measurement of single-copy orthologs that are universally present across diverse species (Waterhouse et al. 2018; Manni et al. 2021). In summary: BUSCO provides quantitative measures of the completeness of a dataset in terms of expected gene content. It assesses the number of complete BUSCOs (those found in their entirety), fragmented BUSCOs (only a piece of them is identified), missing BUSCOs (not identified), and duplicated BUSCOs (found more than once). The BUSCO tool achieves this by searching the dataset with a set of lineage-specific profiles. These profiles are built using hidden Markov models (HMMs), which are statistical models that are able to capture the patterns in a set of sequences. In this case, the sequences are protein sequences from a set of “benchmariking universal single-copy orthologs” – genes that are expected to be found in a single copy in every species of the group under consideration. The choice of lineage will depend on the organism under study. BUSCO results are straightforward to interpret and give a good sense of the quality and completeness of the dataset (Waterhouse et al. 2018; Manni et al. 2021). Using lineage-specific datasets from BUSCO (say which one), I was able to quantitatively evaluate the completeness of the genomic transcriptomic and proteomic databases I previously obtained. This process ensured the reliability of the data by identifying complete, fragmented, duplicated and missing orthologs. Not only did BUSCO assist in determining the overall quality of the data, but it also enabled the identification of potential gaps or duplications within these datasets. This rigorous assessment was essential in ensuring the robustness of the subsequent analyses and findings in my research because it allowed me to select representative species from different families of the tree of life for each project, for example in chapter 3 I used a collection of X sequences for the analysis of of X, for the chapter 5 I only used animal species, including this and that (see table X).

**Phylogenetic analyses**

***Data mining and optimising of final gene family datasets***

With the previously collected queries and species database, we can now perform the first step of phylogenetic analysis which is the data mining. I used one of the most widely used methods for this which is BLAST. There are several different types of BLAST (e.g. if done on amino acid or nucleotidic sequences or if sequence motif information is used) but the basics of it are x,y,z. This first step can serve as a starting point for further refinement of gene families to use for phylogenetic trees. In fact the precision of the results depends on the evalue cut off chosen. The evalue is bla bla.

While the starting point of using BLAST was used as a first step in all of my chapters and is a very common method. Additional methods to refine the gene families can be more varied. Specifically, for my phd I tested two main strategies. One was to filter the results of blast with ad hoc information for each family. Meaning that I filtered out for known protein domains present in each protein family. This is a very precise strategy and should provide high quality results. Another strategy is less targeted but potentially quicker and it is to cluster the sequences to determine the degree of relatedness. This can both serve to determine cut-offs to remove spurious sequences collected with blast but not related enough and it can help to discriminate between different gene families, where the distinction is not that clear a priori (e.g. because they are part of a larger superfamily or because literature describes them as having same function but nothing is known about their evolutionary relationship). In turn this second strategy has different ways to be approached. In chapter 5 a basic tool that clusters sequences based on all-vs-all blast scores was used. While in chapter 4 more sophisticated methods which include the definition of orthogroups were used. In the latter case the first step of these tools is also to cluster based on similarity scores, but then additional approaches are used, e.g., phylogenetic or network analysis.

***Annotating Sequences***

A useful additional step is to provide annotations to the sequences collected. This is because not all species proteomes came with annotations. And to navigate the big trees or clusters of sequences it is useful to have as many sequences as possible with a “name”. The most commonly used method throughout this thesis is to blast sequences vs swissprot and keep the best hit. Where necessary, further manual annotations were applied by taking advantage of species-specific or taxon specific databases. Throughout this thesis useful species or taxon specific databases used were: genecards for human, xx for mouse, flybase for flies, echinobase for sea urchins and other echinoderms and yy for Arabidopsis thaliana.

***Multiple sequence alignment and trimming***

Once final gene families have been optimally curated. The next step is to align the sequences. The idea is that if the sequences are homologous they can be aligned in such a way that each column of the alignment represents a homologous position in a consensus sequence. Where all or most sequences align perfectly there is high conservation of the sequence, where alignment is less good there are more divergent portion of the sequence. And finally there can be gaps throughout the alignament that can either represent true signal or in some cases may be due to technical difficulties. In any case, before moving forward with phylogenetic tree construction, it is useful to remove potentially less useful positions in the alignment. Meaning, if there are too many gaps in a specific position of the alignment this could be problemetic in the tree and contemporarily increase computation times without necessarily adding much phylogenetic information. Conversely some gaps can be useful information. Therefore we use tools that can automatically decide cutoff for trimming alignments. And these can be further tailored according to need. e/g/ if we have short sequences to start with then we don’t want to trim too much if not we are not left with enough phylogenetic signal for the trees.

***Gene tree construction***

Once the msa is ready it can be used for tree construction. All trees in this thesis were built under maximum likelihood. Generally, a model finder was run to determine the best fit tree. Supports for the tree were generally UFB, however, TBE trees were used in chapter 5.

Gene tree to species tree reconciliation

In some cases, it can be useful to re-infer a gene tree on the basis of the knowledge of the species relationships. This can be done using generax.

Single-cell sequencing analyses

For one of my aims – understanding the molecular setup of photoreceptor cells (Chapter 3) I used also single cell sequences analyses of publicly available data. This was because, after determining the presence/absence of phototransduction genes in the genome of target species, I wanted to know whether they were co-expressed within the same cell type – a candidate PRC cell. Furthermore, I wanted to identify other genetic patterns common to animal PRCs. By analysing sc data you can see if there are other genes in common in the cell type. With focus on regulatory genes.

**Preliminary steps**

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.

-downloading of the datasets: in my case what I needed was the gene to cell matrix. Say quickly what must have been the steps that the authors used previously to get to this step.

**Clustering cells into metacells**

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 – so see more details in respective chapter (chapter 3).

References

Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols* 1:e323.

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* 8:eabi5884.

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* 16:28.

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.