Chapter 2

General Methods

Background about general methodologies

To address the broad research questions of my thesis – the evolution of vision and the evolution of chemokine signalling, I used various bioinformatic methodological approaches. While exact methods are described precisely in each 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. Therefore, this chapter provides a comprehensive overview of the central methodologies, which will act as the shared common foundation for the next chapters.

Data acquisition and BUSCO analyses.

The comparative analysis of systems and signalling pathways in different organisms requires of the examination of genomes and predicted proteomes across diverse species. Thus, acquiring high-quality data from different sources and different species was one of the initial steps in my research. For this, I obtained large databases including (how many species I have in my case, describe datasets and how they were obtained).

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).

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).

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.