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 expansive 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 2019 for eukaryotic classification and Burki 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 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 acid 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 and 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 and 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 and 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 and 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 paralogous (resulting from gene duplication) and orthologous (stemming from speciation) sequences. 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 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. 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, 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