ICAR-Indian Institute of Spices Research, Kozhikode, Kerala, India



Training ReportonPost-MSc Training in BioinformaticsConducted atBioinformatics and Integrative Genomics Facility, Division of Crop Improvement and Biotechnology.

Submitted by

Kavya K M

Post-MSc Trainee at ICAR-IISR, Kozhikode

Under the supervision of

Dr. Sona Charles

Scientist

Bioinformatics and Integrative Genomics Facility, Division of Crop Improvement and Biotechnology, ICAR- Indian Institute of Spices Research, Kozhikode- 673012

Duration of Training

[19/08/2024] - [16/06/2025]

Post-MSc Training in

Bioinformatics at ICAR-

IISR: Training Report

# ACKNOWLEDGMENT

Without the support, patience, and guidance of several individuals, this training period would not have been completed. I owe my deepest gratitude to the Bioinformatics and Integrative Genomics Facility, ICAR-Indian Institute of Spices Research, Kozhikode, Kerala, for providing me with this valuable opportunity.

I am especially grateful to Dr. Sona Charles, Scientist at ICAR-IISR, for the chance to undertake this training program under her mentorship. Her guidance has been instrumental in my development throughout this period. I also extend my sincere thanks to Dr. Manimaran, Scientist at ICAR-IISR (Discipline – Nematology), for his valuable support and encouragement during the course of this training.

I extend my sincere and heartfelt thanks to ,Ms. Shahana Arif (Project Associate), Dr. Silju Juby (Young Professional 1), and Mr .M A Fayad (PhD student), for their constant support and guidance. Their insights and encouragement greatly enhanced my learning experience.

This project could not have been completed without the unwavering support of my family, who have always been my pillars of strength and encouragement.

Last but not least, I am profoundly grateful to almighty God for His blessings throughout this journey.

CONTENTS

| **Sl. No.** | **Section Title** | **Page No.** |
| --- | --- | --- |
| 1 | Abstract | 5 |
| 2 | Introduction | 6 |
|  | - Purpose of the Training | 7 |
|  | - Scope of Training | 7 |
| 3 | Workshops and Conferences Attended | 8 |
| 4 | Structure of the Report | 8 |
| 5 | **Part I: Molecular Docking Studies** |  |
|  | - Introduction to Molecular Docking | 10 |
|  | - Clove Oil and Eugenol | 10 |
|  | - Dental Nociceptors | 11 |
|  | - Methodology | 12 |
|  | - Results | 14 |
|  | - Conclusion | 16 |
| 6 | **Part II: Transcriptomics and Comparative Genomics** |  |
|  | - Introduction to Transcriptomics | 17 |
|  | - DEG Analysis | 17 |
|  | - Visualization of DEG (Heatmaps & Volcano Plots) | 18 |
|  | - Orthologous Analysis (OrthoVenn3 and OrthoFinder) | 23 |
|  | - Results Interpretation from OrthoVenn3 | 24 |
|  | - Phylogenetic Tree Construction | 24 |
| 7 | Article Summary: Hi-C Sequencing | 30 |
| 8 | Conclusion | 30 |
| 9 | References | 32 |

**ABSTRACT**

The post-MSc training in bioinformatics at the ICAR-Indian Institute of Spices Research (IISR), Kozhikode, provided a comprehensive and immersive learning experience aimed at bridging theoretical knowledge with practical application. This program was specifically designed to cultivate strong hands-on skills in bioinformatics and deepen my understanding of modern computational approaches in biological research. During the training, I actively participated in a range of activities that included molecular docking of proteins with ligands to assess binding affinities, statistical programming in R for biological data analysis, and the differential expression analysis of genes across different organisms. I also worked on constructing phylogenetic trees, performing gene and protein BLAST analyses, and investigating orthologous gene relationships using tools such as OrthoFinder and OrthoVenn. These activities enabled me to develop practical expertise in widely used bioinformatics tools and pipelines. I enhanced my skills in statistical data interpretation, gained a foundational understanding of machine learning applications in bioinformatics, and acquired hands-on experience in molecular interaction analysis through docking studies. The training not only improved my technical competencies but also broadened my perspective on how computational biology is applied in real research settings. Overall, the training at IISR, Kozhikode, was a transformative and enriching experience. It equipped me with the necessary skills and confidence to tackle future challenges in the field of bioinformatics and contributed significantly to my professional and academic development.

**INTRODUCTION**

The post-MSc training in bioinformatics at ICAR–Indian Institute of Spices Research (IISR), Kozhikode, was thoughtfully structured to provide practical exposure to a wide range of bioinformatics tools and methodologies. This report outlines my learning journey during the training, the diverse projects I undertook, and the valuable skills I developed over the course of the program.

Bioinformatics is a rapidly advancing interdisciplinary domain that combines principles of biology, computer science, mathematics, and information technology to analyze and interpret complex biological data. It has become an essential tool in modern research, particularly in fields such as agriculture, healthcare, environmental science, and biotechnology. The increasing use of high-throughput technologies—such as next-generation sequencing (NGS), microarrays, and proteomics—has resulted in the generation of vast and diverse biological datasets. These datasets, now widely available in public repositories, hold immense potential for uncovering the genetic, molecular, and functional mechanisms that underpin various life processes.

Despite the abundance of data, extracting meaningful information from these large and complex datasets presents a significant challenge. This is where bioinformatics plays a pivotal role. Through the application of computational algorithms, data mining techniques, and statistical modeling, bioinformatics allows researchers to manage, analyze, visualize, and derive insights from biological data efficiently. The field not only facilitates the identification of genes, proteins, and pathways but also supports the development of predictive models and hypotheses for further experimental validation.

This training provided an opportunity to gain hands-on experience with core bioinformatics tools and methodologies, offering practical insights into how computational approaches are applied in biological research. The training curriculum was thoughtfully designed to bridge theoretical knowledge with practical application. Participants were introduced to a wide range of bioinformatics tools and databases, covering key areas such as sequence alignment, genome annotation, phylogenetic analysis, transcriptomics, and structural bioinformatics. Emphasis was placed on the interpretation of biological data derived from high-throughput experiments, using software platforms and programming environments commonly used in the field. By engaging in hands-on sessions and case studies relevant to spice genomics and agricultural research, trainees were able to develop critical skills needed to handle real datasets, perform data-driven analysis, and contribute meaningfully to ongoing research initiatives.

**Purpose of the Training**

The post-MSc bioinformatics training at IISR Kozhikode was primarily aimed at bridging the gap between theoretical education and practical application in the field. The program was carefully structured to equip participants with essential skills for handling and analyzing large-scale biological datasets, executing complex computational analyses, and actively contributing to current research projects. For me, this training marked a significant milestone in my professional journey. It offered valuable exposure to advanced bioinformatics tools and methodologies, while also allowing me to apply them in a real-world, research-driven context. This immersive experience greatly strengthened my technical capabilities, sharpened my analytical mindset, and boosted my confidence as an emerging professional in the field of bioinformatics.

**Scope of Training**

During the course of the training, I actively participated in a wide range of bioinformatics tasks and gained hands-on experience with several advanced tools and platforms. I performed molecular docking analyses using AutoDock, enabling the exploration of protein-ligand interactions and binding affinities. In the area of transcriptomics, I conducted differential gene expression analysis using tools such as DESeq2, NOISeq, and edgeR, followed by data visualization through heatmaps and volcano plots to interpret the expression profiles. For comparative genomics, I carried out orthologous gene analysis using both OrthoVenn3 and OrthoFinder, which allowed me to identify conserved genes across species. Additionally, I constructed phylogenetic trees to examine evolutionary relationships among orthologs. These activities collectively enhanced my technical expertise and deepened my understanding of computational approaches in functional and comparative genomics.

**Workshops and conferences attended**

* Workshop on R for Bioinformatics, conducted by ICAR-Indian Institute Of Spices Research during 4th-8th November 2024
* Computational Mining and Integrative Omics for Biomedical Research, MOTIF 2025, organized by Department of Bioinformatics, Bharathiyar University, Coimbatore held on 14.02.2025
* National Symposium on Spices and Aromatic crops, SYMSAC-XI conducted by ICAR-IISR during 7-9 January 2025
* National Symposium on Recent Trends in Omics in Plant Biology and presymposium workshop on Genome wide Association Studies and Transcriptomics: from Reads to Pathways organized by ICAR-Indian Institute Of Spices Research during 4th-8th November 2024 on 20th May.

**Structure of the Report**

This report is based on two distinct yet complementary studies that explore the applications of bioinformatics in both structural and functional analysis.

The first part focuses on a molecular docking study aimed at evaluating the potential of clove oil as a natural remedy for dental pain. Specifically, we investigated the binding interaction between eugenol, the primary active compound in clove oil, and dental nociceptor proteins associated with pain perception. Using computational docking techniques, we analyzed the affinity and binding modes of eugenol to assess its potential efficacy in modulating pain-related pathways at the molecular level.

The second part presents a transcriptomic and comparative genomics analysis, where high-throughput sequencing data were utilized to explore gene expression patterns and evolutionary relationships. This segment involved differential gene expression analysis to identify functionally relevant genes, followed by ortholog identification and phylogenetic tree construction to compare genomic features. Together, these studies demonstrate the diverse applications of bioinformatics—from structural interaction analysis to large-scale gene expression and evolutionary investigations.

Together, these investigations underscore the integrative power of bioinformatics in addressing diverse biological questions—from molecular interactions at the protein level to systems-wide gene expression and evolutionary patterns. By applying both structure-based and function-based approaches, this report highlights how computational tools can contribute to drug discovery, natural product validation, and deeper genomic understanding. These complementary studies not only enhanced my technical competencies but also reinforced the value of interdisciplinary strategies in modern biological research.

1. **Molecular docking studies**

Molecular docking is a computational method used to predict the interaction between two molecules—typically a small molecule (ligand) and a larger macromolecule (receptor or protein target). The fundamental goal of docking is to predict the optimal binding orientation and estimate the strength of interaction between the ligand and the receptor. This is achieved through algorithms that simulate how a ligand fits into the binding site of the target protein, mimicking the key molecular recognition events that occur in biological systems.

The docking process generally involves two critical components:

* Search Algorithm – which explores all possible conformations and orientations (poses) of the ligand within the binding site of the protein.
* Scoring Function – which evaluates and ranks these poses based on predicted binding affinity, taking into account various interactions such as hydrogen bonding, hydrophobic interactions, van der Waals forces, and electrostatic complementarity.

The protein-ligand complex is often treated as a lock-and-key or induced fit model, where either the ligand fits precisely into a pre-defined active site or both the ligand and receptor undergo slight conformational changes upon binding. Docking algorithms may also incorporate flexibility in either the ligand, the receptor, or both to better mimic real biological environments.

Molecular docking is widely used in drug discovery, natural product screening, and structure-based functional analysis, as it enables rapid virtual screening of large compound libraries to identify potential lead molecules before investing in costly laboratory experiments. In the context of natural products like eugenol from clove oil, docking provides insights into how such compounds may exert therapeutic effects by binding to biological targets involved in disease mechanisms.

In this study, docking was employed to investigate the interaction between eugenol and various dental nociceptor proteins, providing a predictive model for its potential analgesic effects in dental pain relief.

**Clove Oil and Eugenol**

Clove (Syzygium aromaticum) is a medicinal plant long used in traditional and modern dentistry for its potent analgesic, anti-inflammatory, and antimicrobial properties. It is especially effective in managing dental pain, making it a widely used remedy in oral healthcare. The primary active constituent of clove oil is eugenol, which comprises approximately 70–85% of the essential oil (Cortés-Rojas et al., 2014). Eugenol exhibits anesthetic and anti-nociceptive effects by acting on peripheral pain receptors, particularly nociceptors involved in toothache and other oral pain. It functions by modulating ion channels and inhibiting prostaglandin synthesis, thereby blocking the transmission of pain signals.

In addition to eugenol, clove contains several other bioactive constituents such as β-caryophyllene, acetyl eugenol, vanillin, tannins, and flavonoids, which contribute synergistically to its therapeutic profile. Due to its ability to relieve dental pain naturally, clove and its derivatives are widely included in dental formulations such as medicated toothpastes, mouthwashes, and temporary fillings.

In this study, eugenol is used as the principal ligand for molecular docking analysis to investigate its interaction with nociceptor receptors implicated in dental pain. Its dominant presence in clove oil and its well-documented pharmacological activity make it an ideal candidate for structure-based virtual screening.

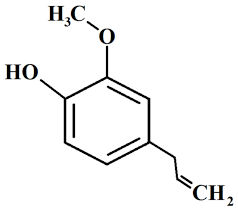


Fig 1: Chemical Structure Of Eugenol

**Dental Nociceptors**

Dental nociceptors are specialized sensory receptors responsible for detecting noxious stimuli such as mechanical pressure, thermal fluctuations, and chemical irritants in the oral cavity. These receptors play a critical role in the perception of dental pain, acting as the primary transducers of painful signals from the tooth pulp and surrounding tissues to the central nervous system. Nociceptors include a variety of ion channels, receptors, and transmembrane proteins, such as TRPV1, ASIC3, P2X3, and Nav1.7, which respond to different pain-inducing stimuli.

For this study, a focused literature review was conducted to identify key receptors associated with dental nociception. Based on the findings, 19 relevant dental nociceptor proteins were selected due to their documented involvement in pain signaling pathways. These receptors represent major molecular targets through which analgesic compounds like eugenol may exert their effects. The three-dimensional structures of these proteins were retrieved from the Protein Data Bank (PDB) in PDB format, to be used as receptors in molecular docking simulations. This curated selection enabled a comprehensive evaluation of eugenol's potential to interact with multiple targets implicated in dental pain.

**Methodology**

The current study was designed to investigate the interaction between eugenol, the primary bioactive constituent of clove oil, and dental nociceptor proteins implicated in pain perception. The methodology adopted for this study comprises the following steps:

Selection and Retrieval of Receptors:

A comprehensive literature review was conducted to identify proteins associated with dental pain. Based on their reported involvement in nociception, a total of 19 dental nociceptor receptors were selected. The three-dimensional structures of these proteins were retrieved in .pdb format from the Protein Data Bank (PDB) (Berman et al., 2000).

Binding Site Identification:

To predict the potential ligand-binding pockets on the receptor proteins, P2Rank was used. It is a machine learning-based ligand-binding site predictor that provides fast and accurate results without requiring complex input (Krivák & Hoksza, 2018). This step was crucial for defining the grid box parameters required in molecular docking.

Ligand Preparation and Docking Studies:

The ligand, eugenol, was selected as the principal compound from clove oil for its known analgesic properties. Its structure was retrieved from PubChem in .sdf format and converted to .pdbqt format using Open Babel (O'Boyle et al., 2011). Docking simulations were carried out using AutoDock 4.2 (Morris et al., 2009), where eugenol was docked into the active sites of each selected receptor to evaluate their binding affinity and interaction pattern.

Interaction Analysis:

The resulting docked complexes were analyzed using a suite of tools to assess protein-ligand interactions:

PLIP (Protein-Ligand Interaction Profiler) (Adasme et al., 2021), for automatic detection of hydrogen bonds, hydrophobic contacts, and other interactions.

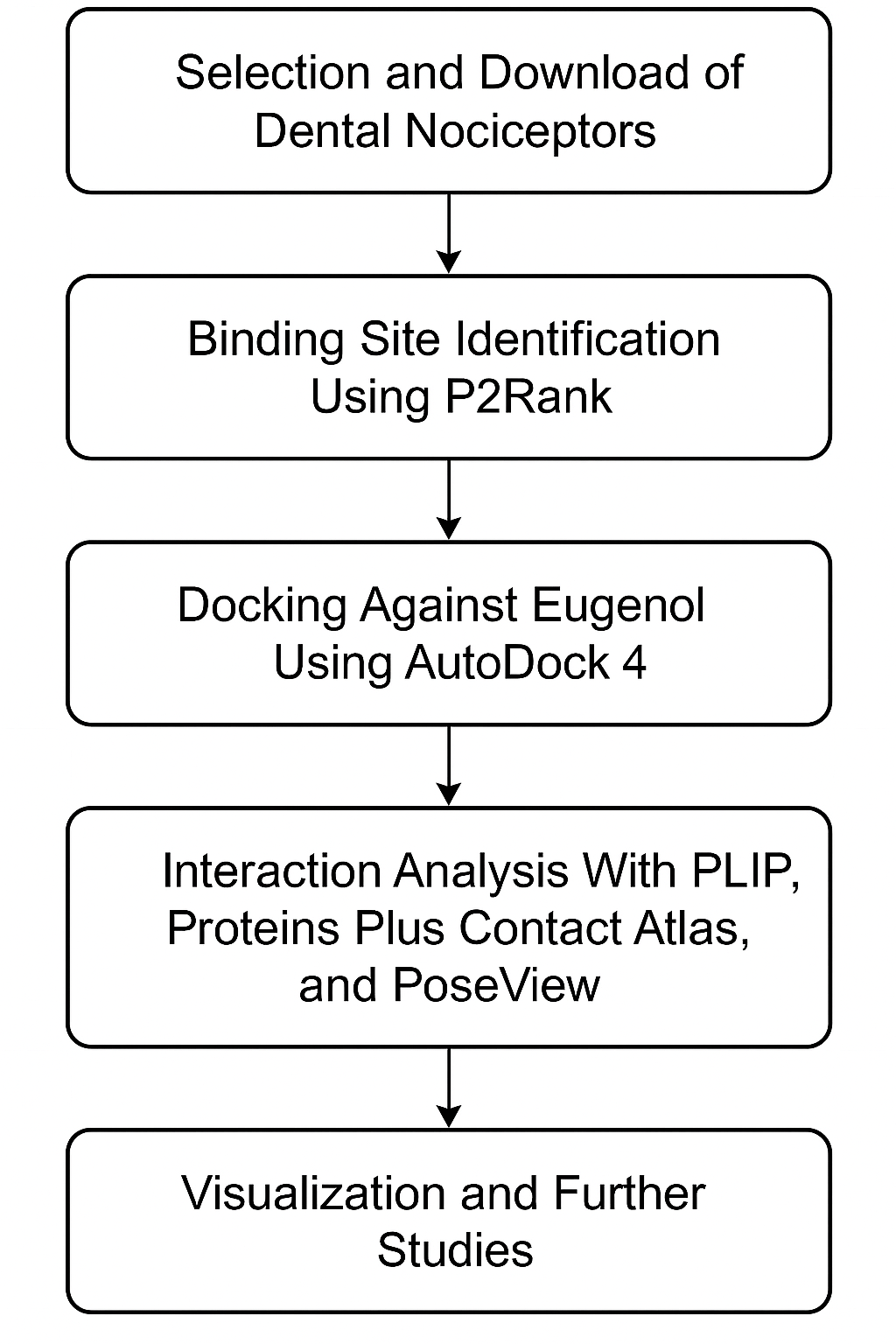
ProteinsPlus Contact Atlas, for advanced interaction mapping and residue-level details.

PoseView, for generating detailed 2D interaction diagrams to visualize bonding interactions clearly.

Visualization and Further Analysis:

The docking results were visualized using PyMOL and Discovery Studio Visualizer, enabling high-quality 3D and 2D representations of ligand-receptor interactions. The binding energies, interaction types, and key residues were compared to determine the most favorable eugenol–receptor complexes.

This multi-step methodology facilitated the in silico evaluation of eugenol as a potential natural therapeutic for dental pain, offering valuable insights into its interactions with molecular targets involved in nociception.

****

**Figure 2: Major steps**

**Results**

Molecular docking of eugenol, the primary bioactive constituent of clove oil, against 19 selected dental nociceptor receptors yielded varying binding affinities and interaction profiles, indicating differential potential for pain modulation.

Among the receptors analyzed, TRPV4 (binding energy: -5.28 kcal/mol), P2X4 (-5.39 kcal/mol), and ASIC3 (-5.85 kcal/mol) demonstrated the highest binding affinities, suggesting strong interaction potential with eugenol. These results point toward their relevance as key molecular targets mediating the analgesic effects of clove oil in dental applications.

In terms of total interactions, TREK2 exhibited the highest number (10 interactions), comprising 7 hydrophobic and 3 hydrogen bonds, followed closely by TRPC6 and TRPV4 with 8 interactions each. Notably, TRPV4, which had a high binding affinity and multiple interaction types (6 hydrophobic, 2 hydrogen bonds), stands out as a promising receptor for further experimental validation.

The diversity in receptor–ligand interaction profiles is indicative of eugenol’s multi-target binding capability, with hydrophobic interactions being the dominant type across all receptors, supported in some cases by hydrogen bonds and π-interactions (e.g., in TRPV1, TRPV2, TRPV4).

The strong binding affinities observed for ASIC3, P2X4, and TRPV4, which are well-established pain-related ion channels, support the hypothesis that eugenol can modulate nociceptive signaling pathways, thereby contributing to its analgesic properties in dental care. These findings align with the traditional use of clove oil for toothache and provide a computational rationale for its therapeutic efficacy.

Overall, the docking results underscore the potential of eugenol as a natural analgesic, targeting key pain receptors, and warrant further experimental validation through in vitro and in vivo studies.

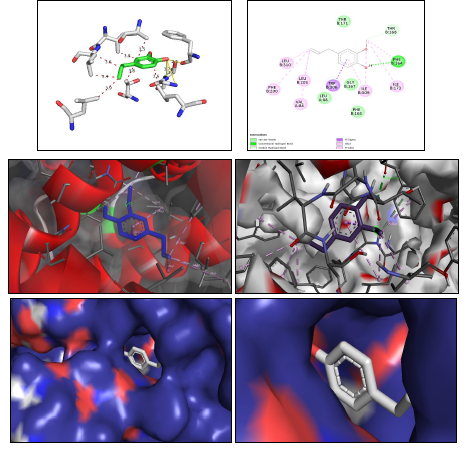
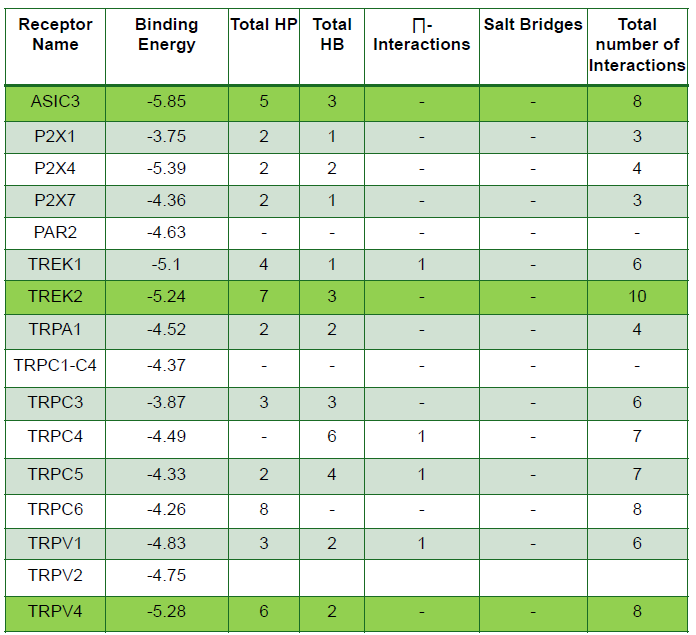


Figure 3: Interaction Diagram of TREK2 receptor and Eugenol

Table 1: Interaction Table showing results of molecular Docking

****

**Conclusion**

This study provides compelling computational evidence supporting the analgesic potential of eugenol, the principal component of clove oil, through its interactions with key dental nociceptor receptors. Molecular docking revealed strong binding affinities particularly with ASIC3, P2X4, and TRPV4, indicating their significance as potential targets for dental pain relief. The diversity and strength of interactions, especially hydrophobic and hydrogen bonds, highlight eugenol’s ability to engage multiple receptor pathways involved in nociception. These findings validate the traditional use of clove oil in dental care and pave the way for further experimental and clinical research to develop plant-based therapeutic alternatives for dental pain management.

1. **Transcriptomics and Comparative Genomics**

Transcriptomics is the comprehensive analysis of the complete set of RNA transcripts (the transcriptome) produced by the genome in a particular cell, tissue, or organism at a specific time or condition. It provides critical insights into gene expression dynamics, enabling researchers to determine which genes are active, their expression levels, and how they respond to various biological states, such as disease, stress, or developmental stages. Unlike genomics, which focuses on the static DNA sequence, transcriptomics reveals the functional output of the genome and reflects the real-time biological activity within cells.

The advent of RNA sequencing (RNA-seq) has revolutionized transcriptomic studies by allowing high-throughput, sensitive, and accurate quantification of transcript levels (Wang et al., 2009). RNA-seq surpasses earlier techniques like microarrays by detecting low-abundance transcripts, novel isoforms, and non-coding RNAs with higher precision.

Commonly used bioinformatics tools and platforms for transcriptomic analysis include:

* FastQC for quality control of raw sequence reads
* Trimmomatic or Cutadapt for adapter trimming and filtering
* STAR and HISAT2 for aligning reads to a reference genome
* FeatureCounts or HTSeq for read quantification
* DESeq2, edgeR, and NOISeq for identifying differentially expressed genes
* ClusterProfiler and DAVID for functional enrichment and gene ontology analysis
* PCA plots, volcano plots, and heatmaps for data visualization

Transcriptomics plays a vital role in various research areas including cancer biology, plant stress responses, microbial studies, and developmental biology. It is also foundational in comparative transcriptomics, where gene expression profiles are compared across species or treatments to identify conserved and divergent regulatory mechanisms (Conesa et al., 2016).

**DEG Analysis**

For the transcriptomic analysis, a diverse set of 21 species spanning multiple taxonomic groups—including insects, plants, algae, arachnids, tardigrades, and bacteria—were selected to examine gene expression profiles across a broad evolutionary spectrum. These organisms represent varied ecological niches and physiological adaptations, making the dataset particularly rich for comparative transcriptomics and functional inference.

Preprocessed transcriptome data and quantification results were already available, allowing us to focus directly on downstream differential gene expression (DGE) analysis. To determine the genes that were significantly upregulated or downregulated under specific conditions or treatments, we evaluated several popular RNA-seq DGE analysis tools, including DESeq2, NOISeq, and edgeR—each based on distinct statistical models and normalization strategies.

After preliminary benchmarking and visual inspections (such as MA plots, volcano plots, and clustering analyses), edgeR was selected as the primary tool for the analysis due to its robust performance in datasets with variable dispersion and small sample sizes (Robinson et al., 2010). All subsequent analyses were carried out using R scripts tailored to handle the multi-species data and to facilitate batch processing of large datasets.

The edgeR workflow involved:

* Creating DGEList objects from count matrices
* Filtering lowly expressed genes
* Normalizing data using TMM (Trimmed Mean of M-values)
* Estimating dispersion
* Fitting the negative binomial model
* Conducting exact tests or GLM-based comparisons
* Extracting significantly differentially expressed genes (DEGs) based on adjusted p-values (FDR < 0.05) and fold change thresholds

This method allowed for a consistent and reliable identification of gene expression changes across the selected species, laying the groundwork for comparative functional enrichment and evolutionary insights.

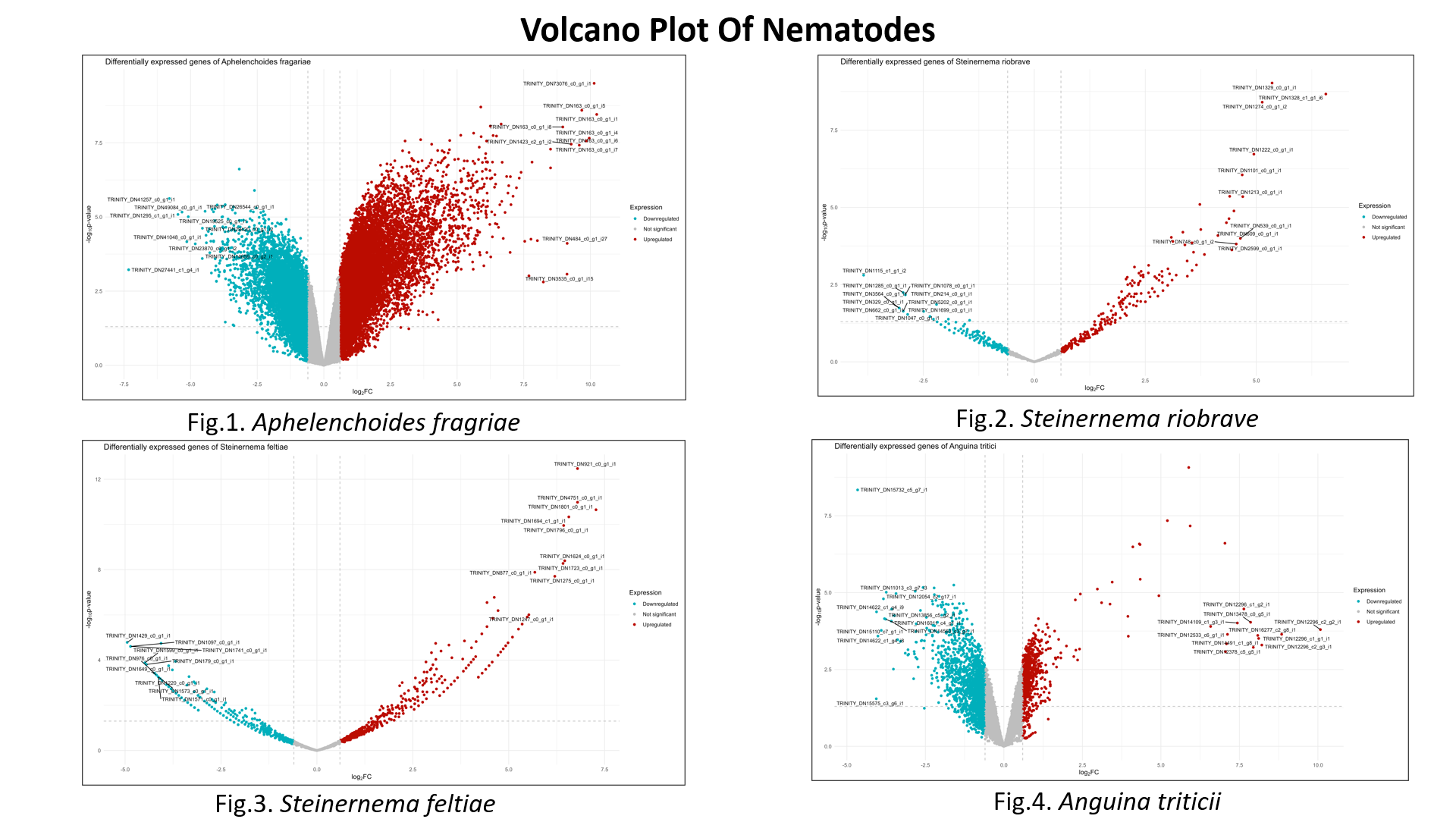
**Visualisation Of DEG**

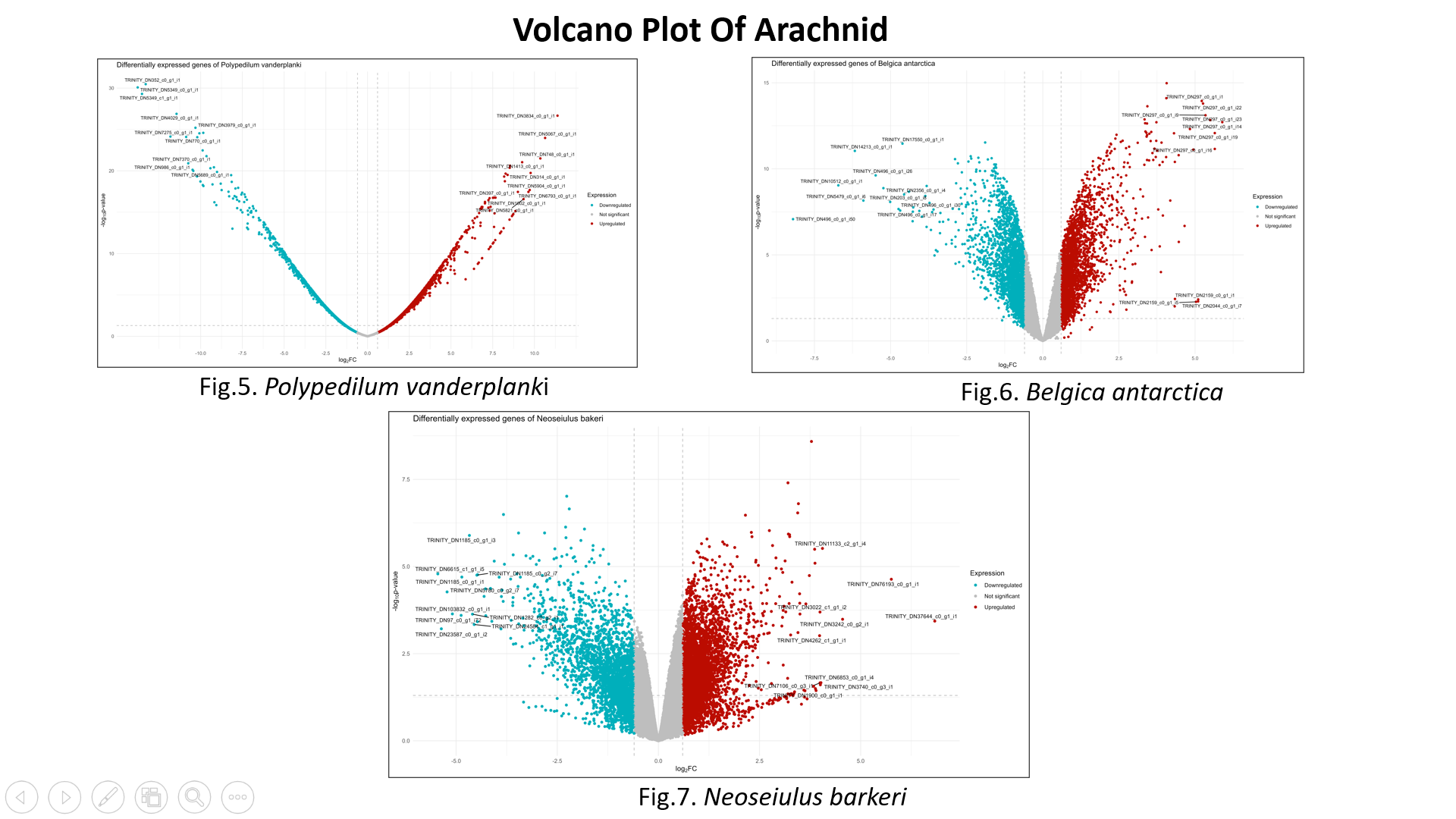
From the differential expression analysis conducted using the edgeR package, genes showing statistically significant changes in expression levels were identified based on adjusted p-values (FDR < 0.05) and log₂ fold change thresholds. To better visualize and interpret the biological relevance of these results, specific visualization techniques were employed using R-based packages.

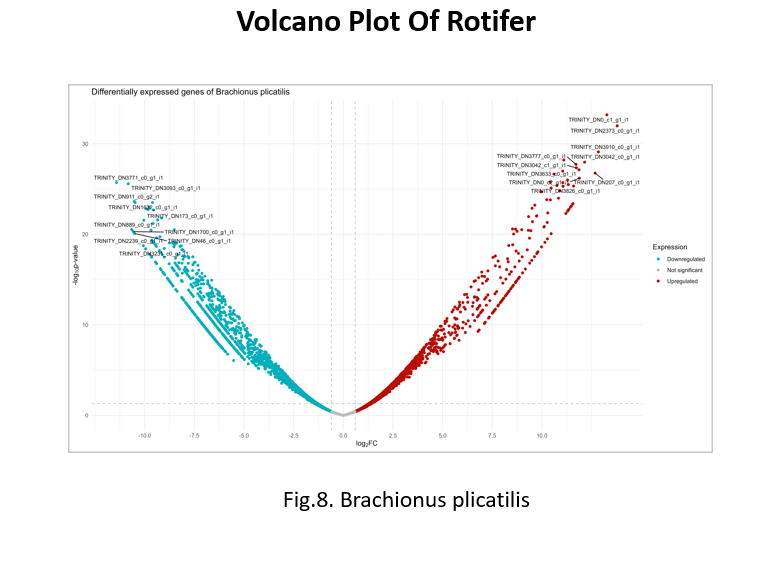
For heatmap generation, the top 10 upregulated and top 10 downregulated genes from each dataset were selected based on their fold change and statistical significance. The pheatmap and ComplexHeatmap packages in R were used to create informative heatmaps that display the expression profiles of these genes across different samples or conditions. These heatmaps helped highlight clustering patterns, group-specific expression, and potential biomarkers among the most responsive genes.

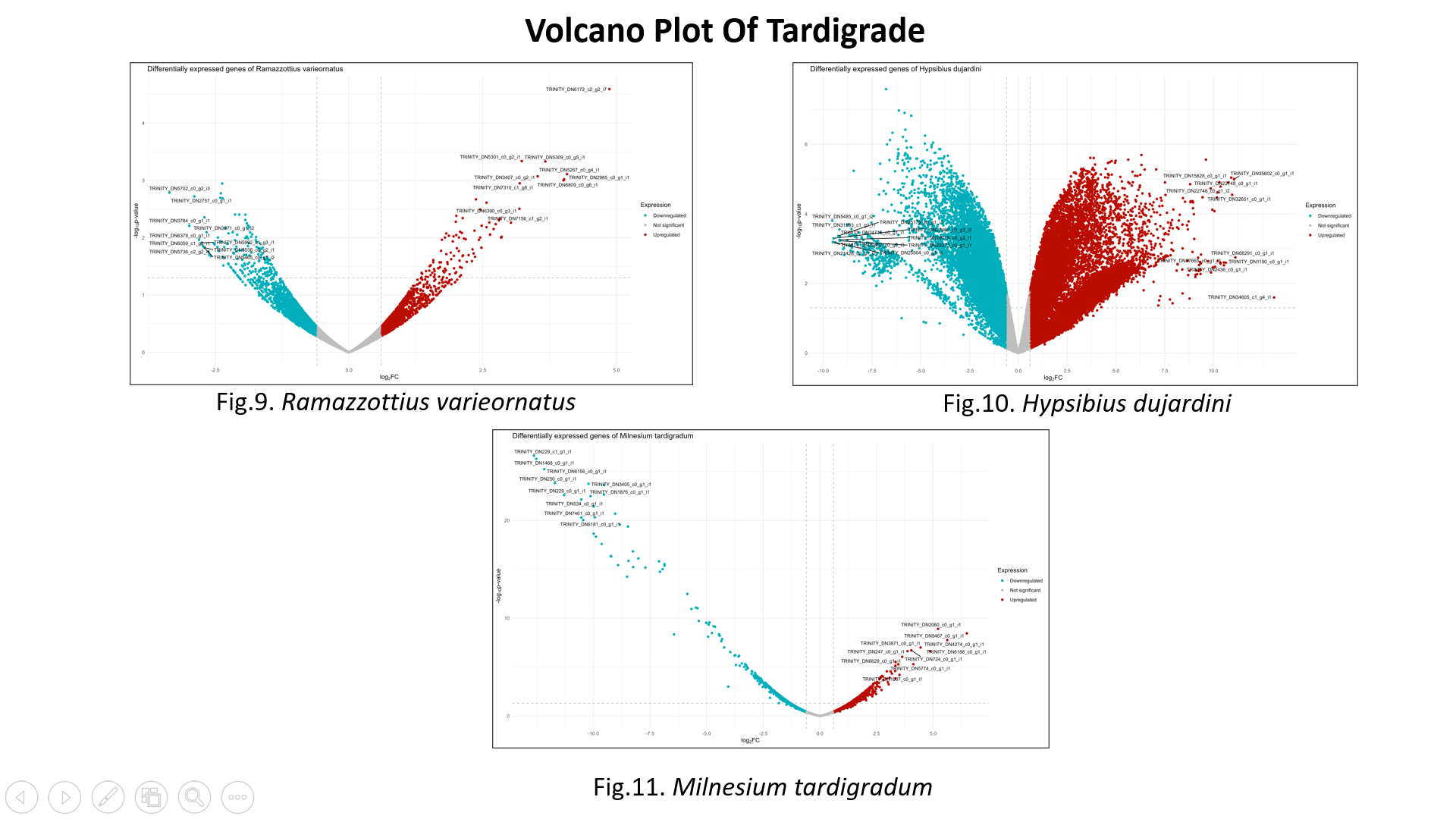
Additionally, volcano plots were constructed to provide a comprehensive overview of gene expression changes by plotting log₂ fold change values against the –log₁₀ adjusted p-values. The EnhancedVolcano and ggplot2 R packages were employed to produce high-quality volcano plots, allowing quick identification of significantly altered genes and visual separation of highly expressed candidates.

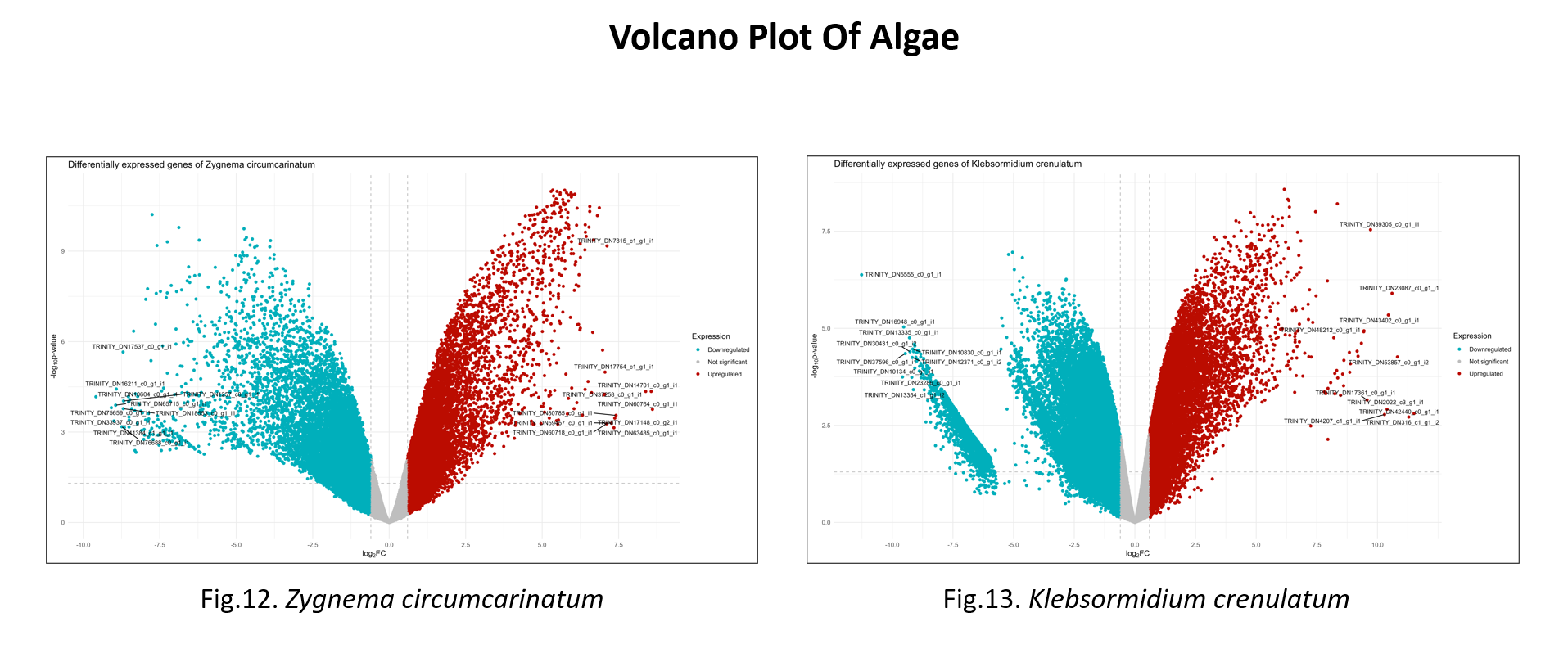
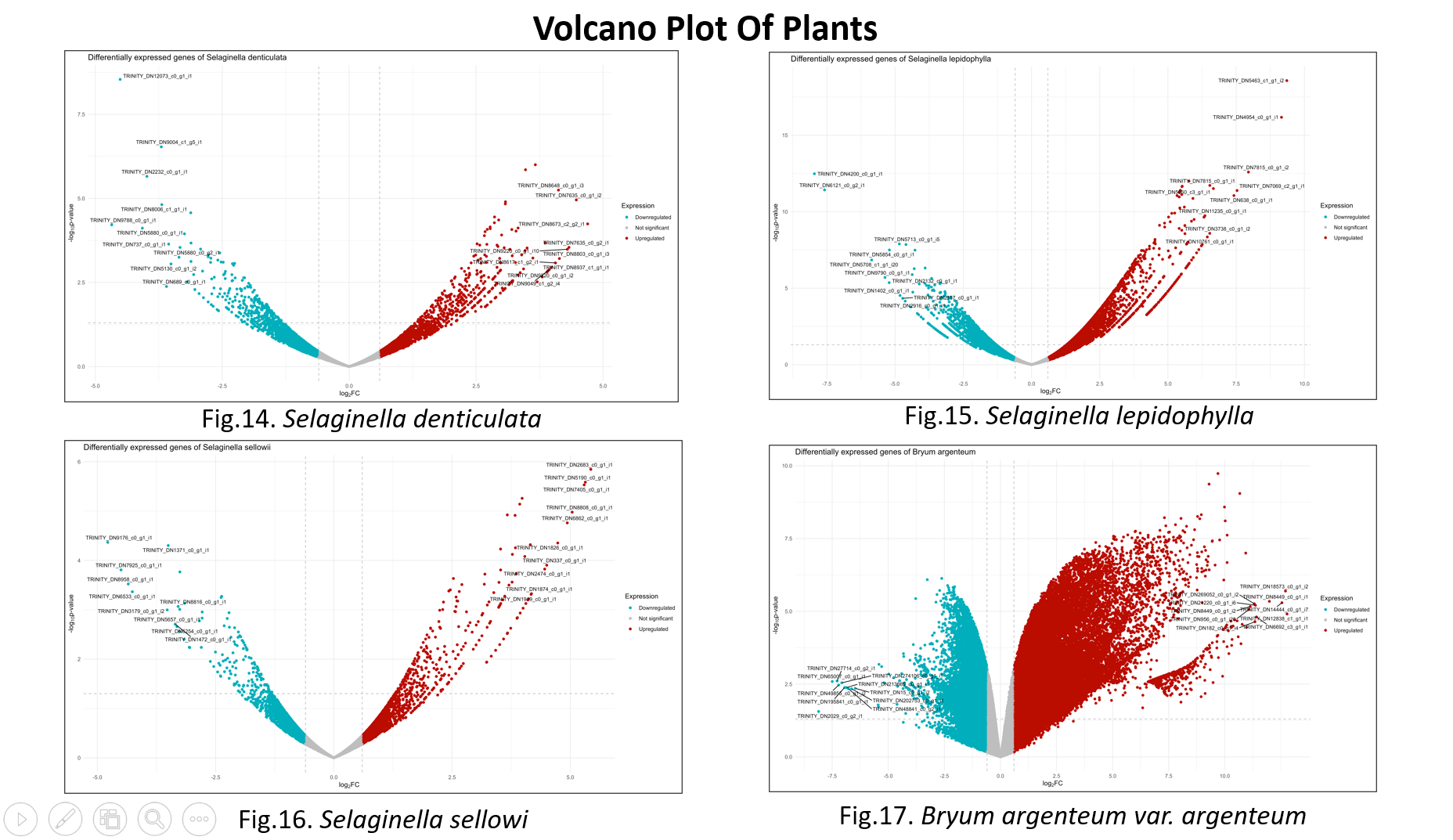
These visualizations not only enhanced the interpretability of the transcriptomic data but also facilitated the selection of biologically relevant genes for further functional annotation and enrichment analysis.

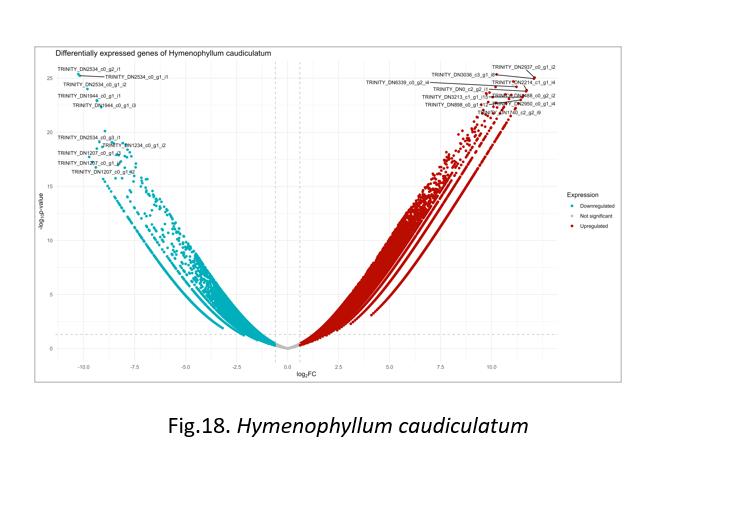


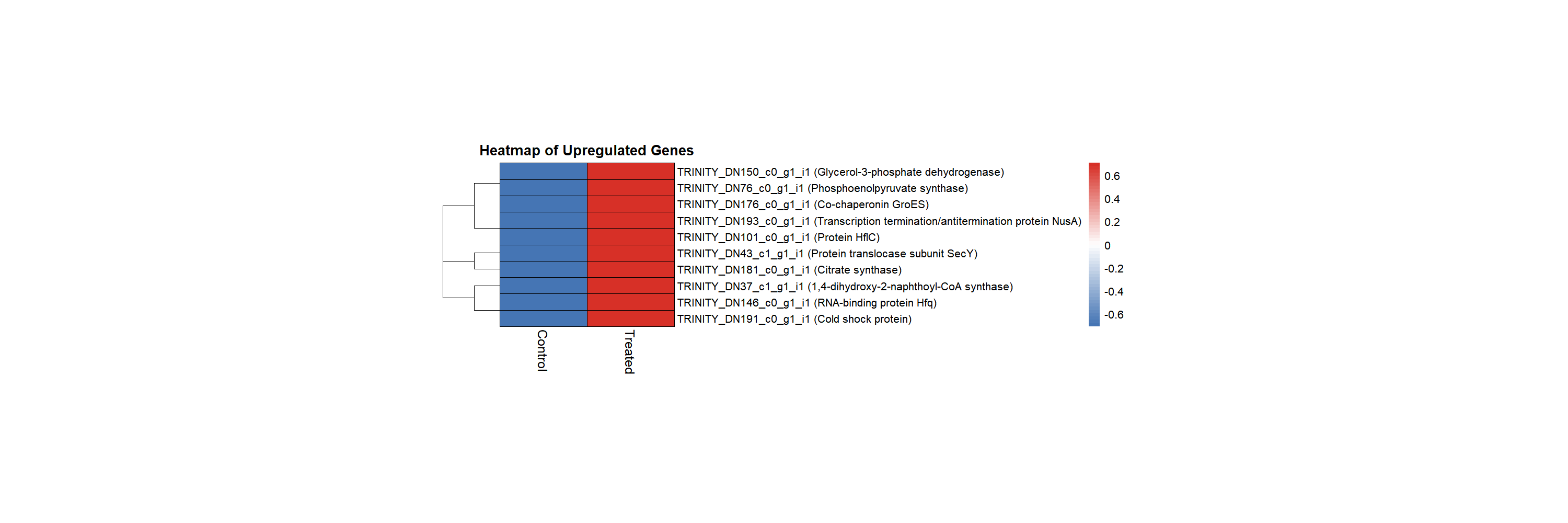
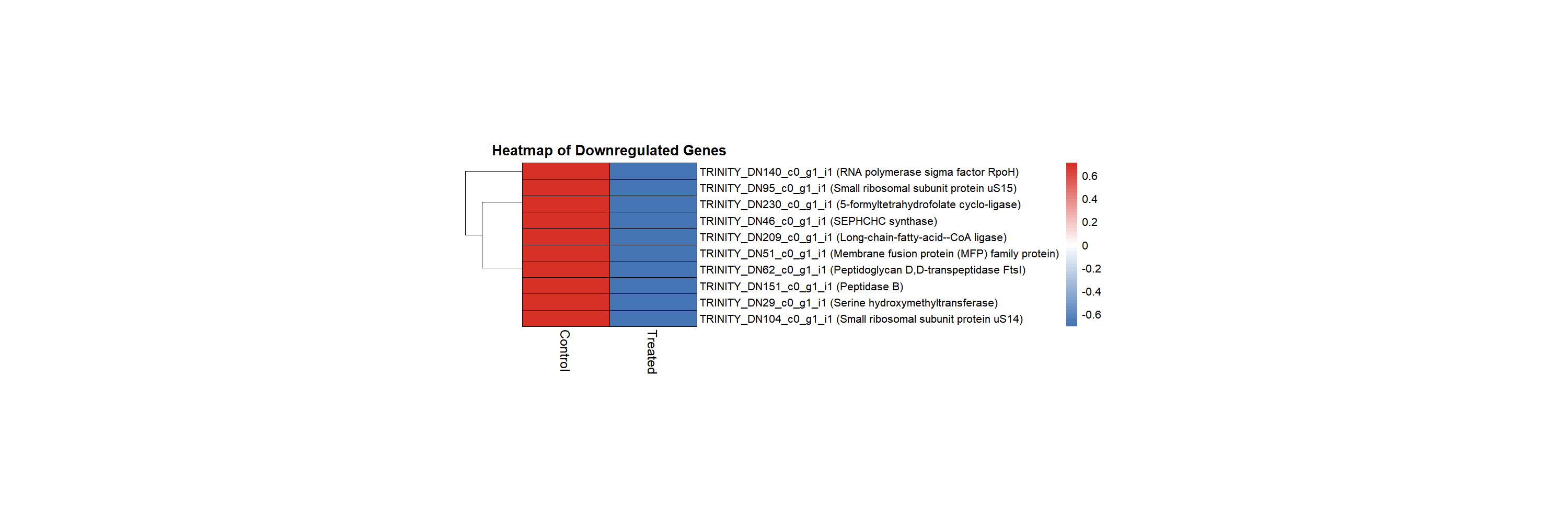








**Heatmap of Photorhabdus luminescens** 

**Orthologous Analysis**

Following the transcriptomic analysis, an orthologous gene analysis was conducted to identify conserved and functionally related genes across the 21 selected species. This step aimed to explore the evolutionary relationships, gene family expansions/contractions, and functional conservation among diverse organisms including insects, plants, algae, arachnids, tardigrades, and bacteria. Two powerful tools were employed for this comparative genomic analysis: OrthoVenn3 and OrthoFinder.

OrthoVenn3 is a web-based platform that facilitates rapid identification and visualization of orthologous gene clusters across multiple species using high-performance alignment algorithms. It offers a user-friendly interface to compare gene families, visualize shared and unique clusters via Venn diagrams, and perform functional enrichment using GO and KEGG databases. OrthoVenn3 uses a graph-based clustering method built on DIAMOND and OrthoMCL, making it ideal for handling large-scale data with speed and accuracy (Xu et al., 2019).

In parallel, OrthoFinder was used as a standalone, high-throughput tool for orthology inference. It is recognized for its robust and accurate phylogenetic framework, capable of producing gene trees, species trees, and inferring orthogroups with high precision. OrthoFinder uses sequence similarity searches (via DIAMOND), followed by clustering with MCL and tree-based inference to distinguish orthologs from paralogs. Its output includes not only orthogroup assignments but also comprehensive phylogenetic relationships and duplication events across species (Emms & Kelly, 2019).

Together, these analyses provided valuable insights into the conserved core genes across taxa, as well as species-specific gene innovations. The identification of orthologous genes laid the foundation for downstream functional annotation, phylogenetic reconstruction, and evolutionary biology investigations.

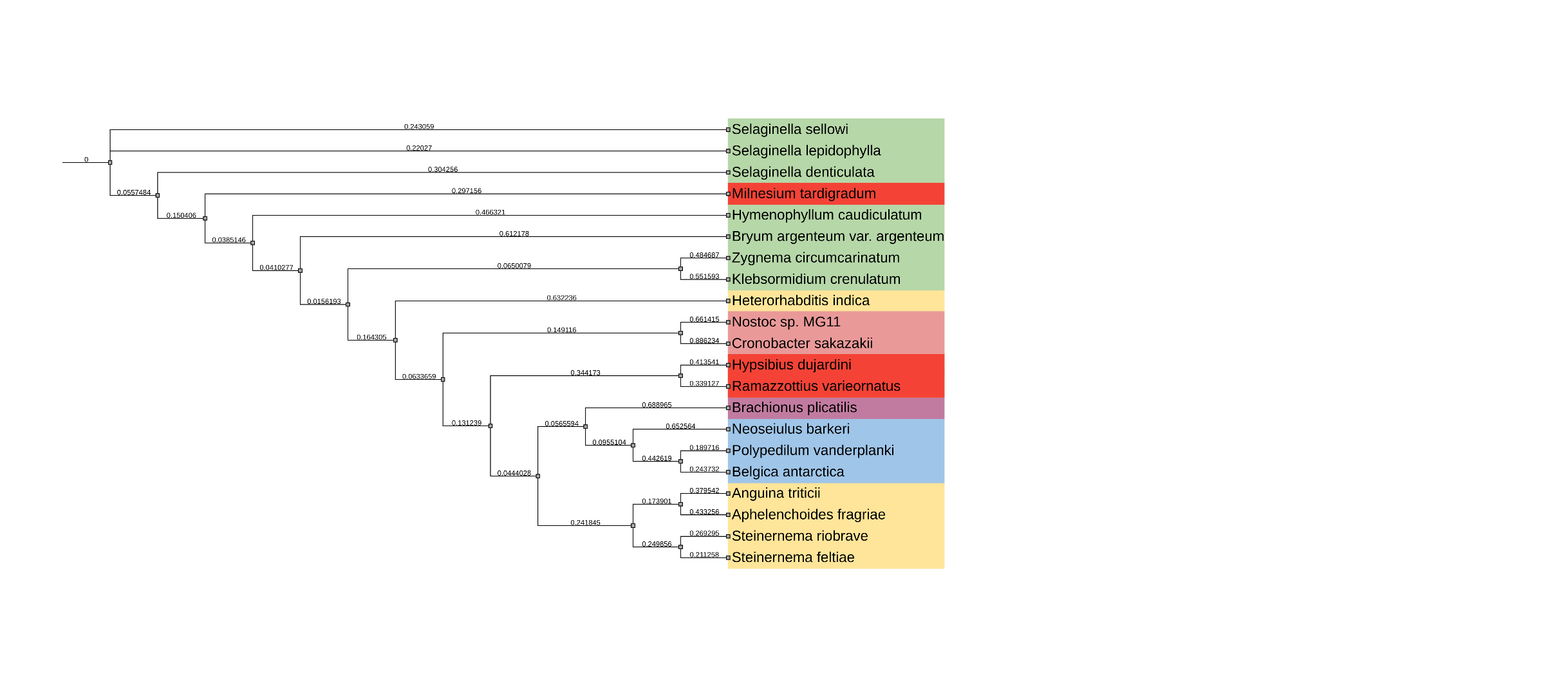


Fig 6: Phylogenetic Tree showing evolutionary relation ship between species

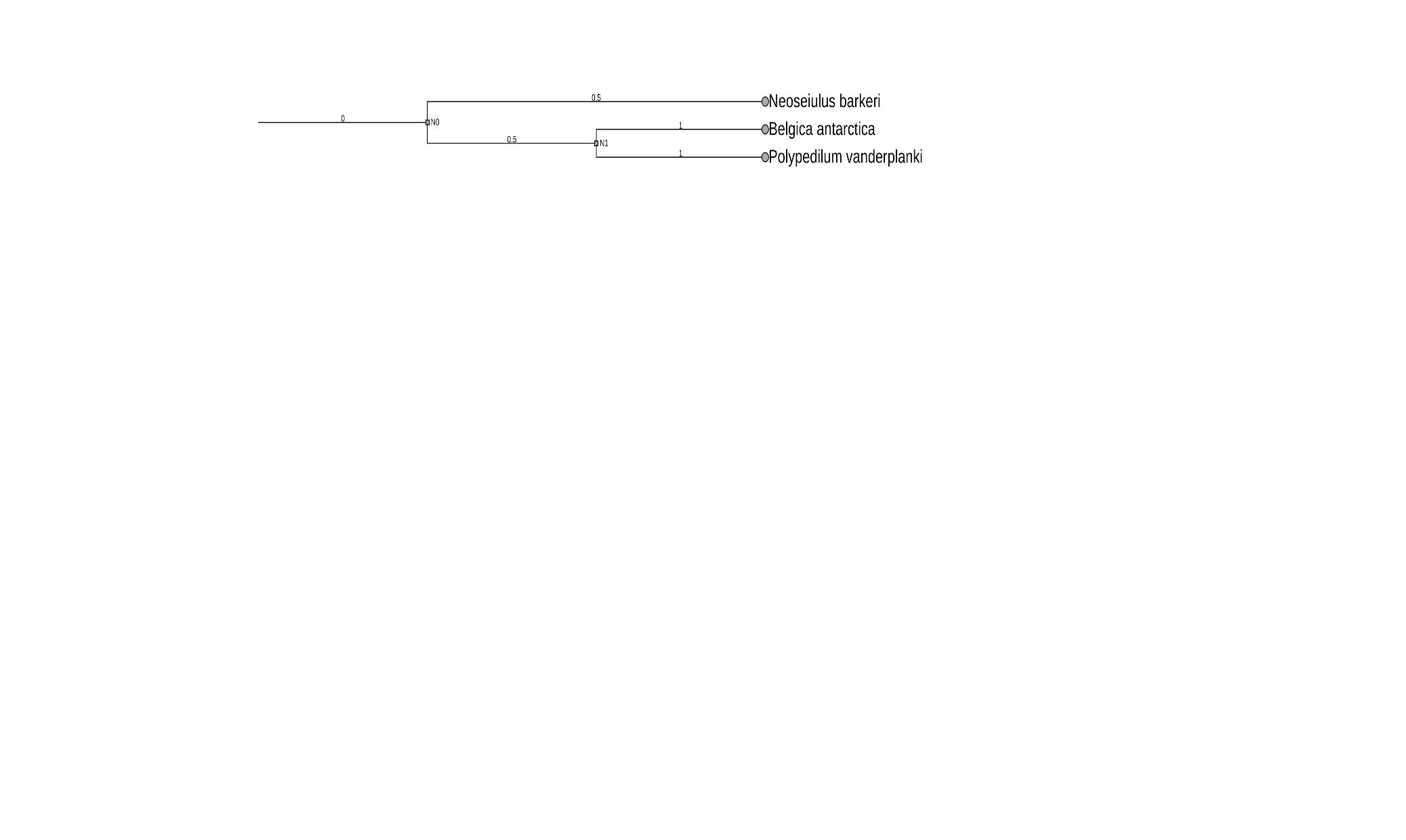
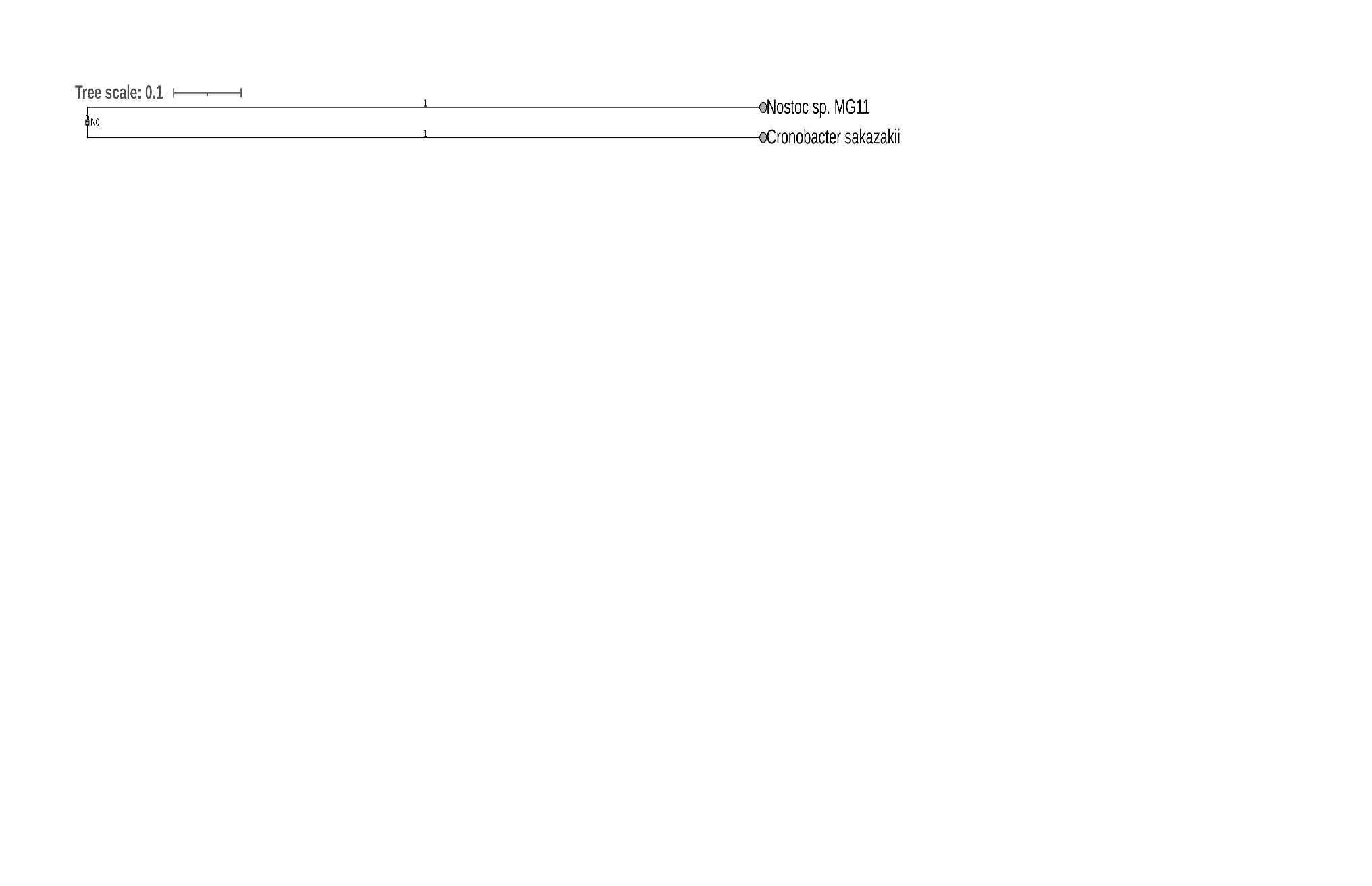
****

Fig 7: Phylogenetic tree of Arachnids

**** Fig 8: Phylogenetic tree of Bacteria

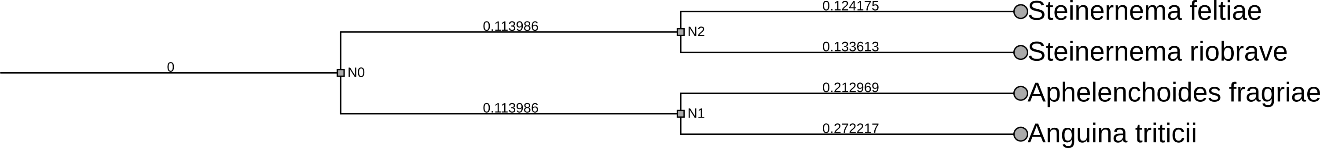


Fig 9: Phylogenetic tree of Nematodes

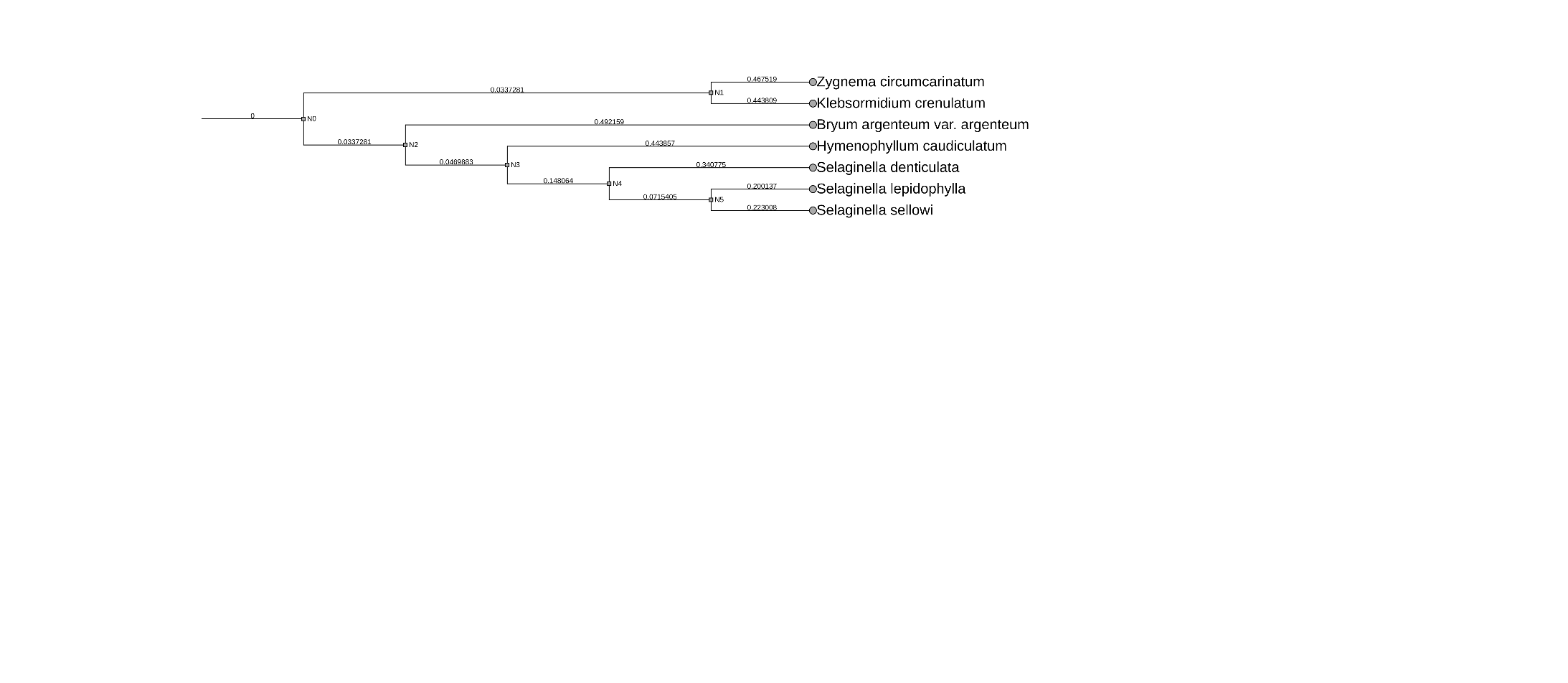


Fig 10: Phylogenetic tree of Plants

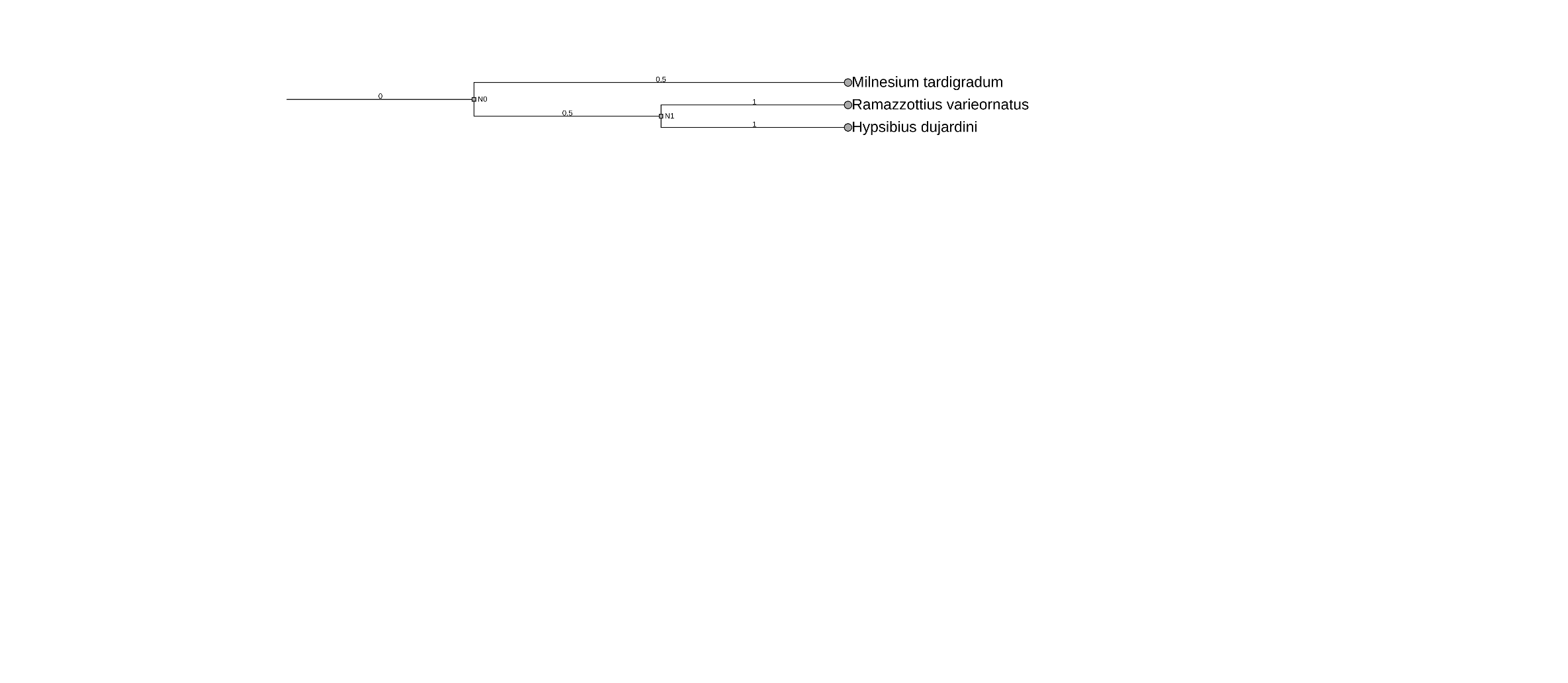
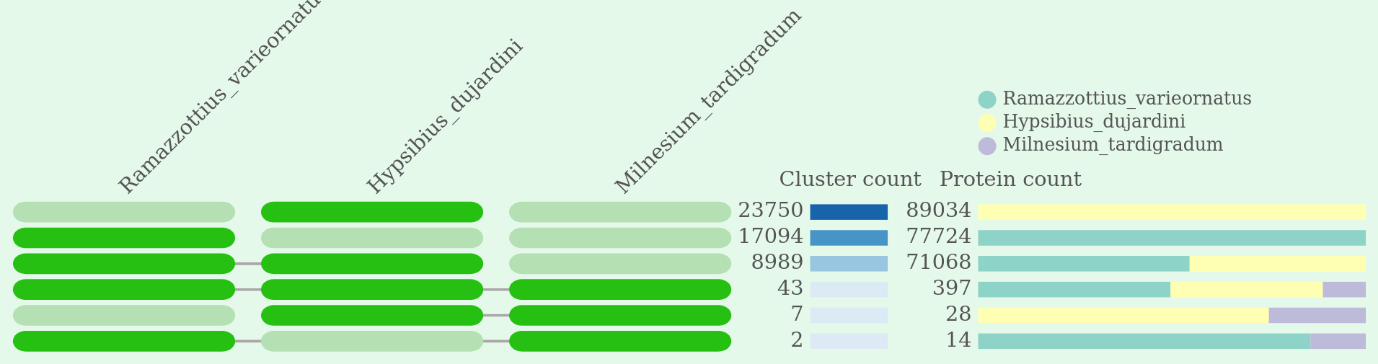
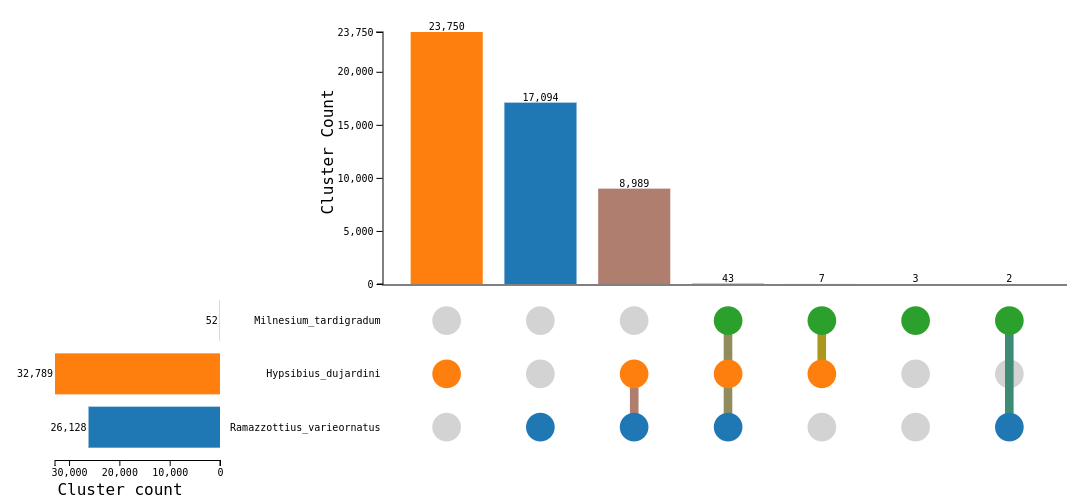
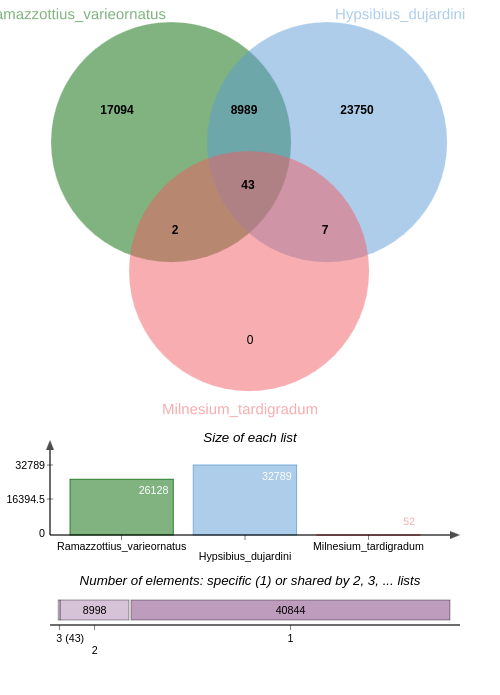
****

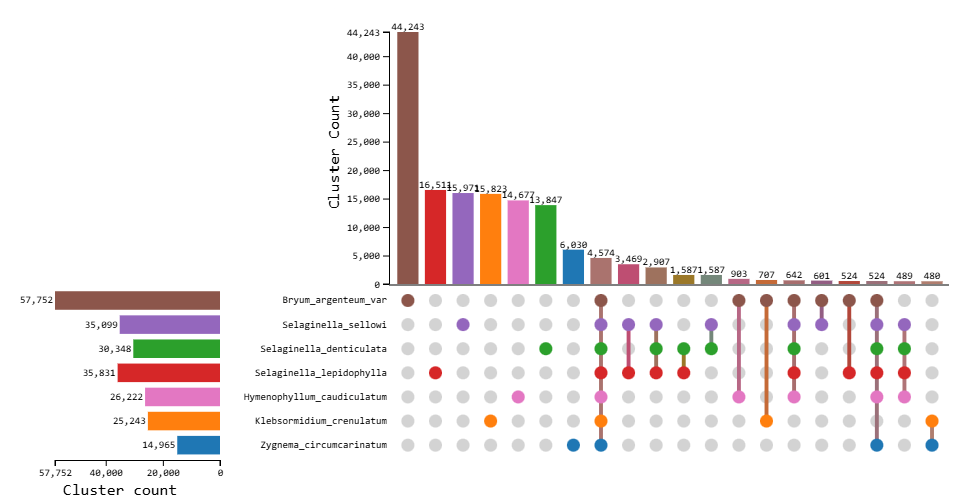
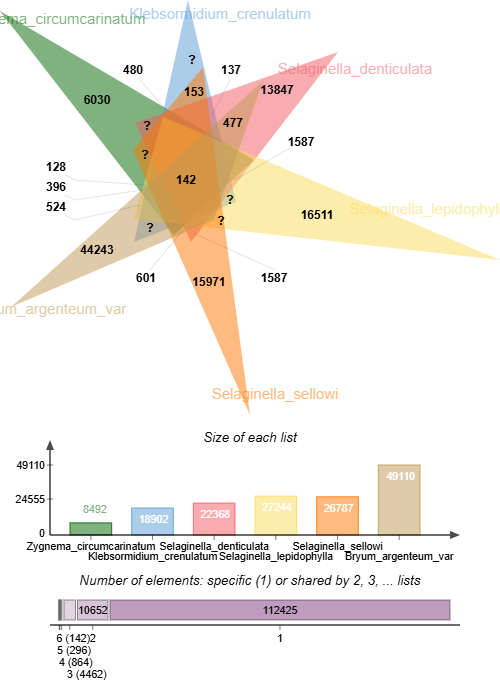
Fig 10: Phylogenetic tree of Tardigrade

**Results from Orthovenn3 for Tardigrades**

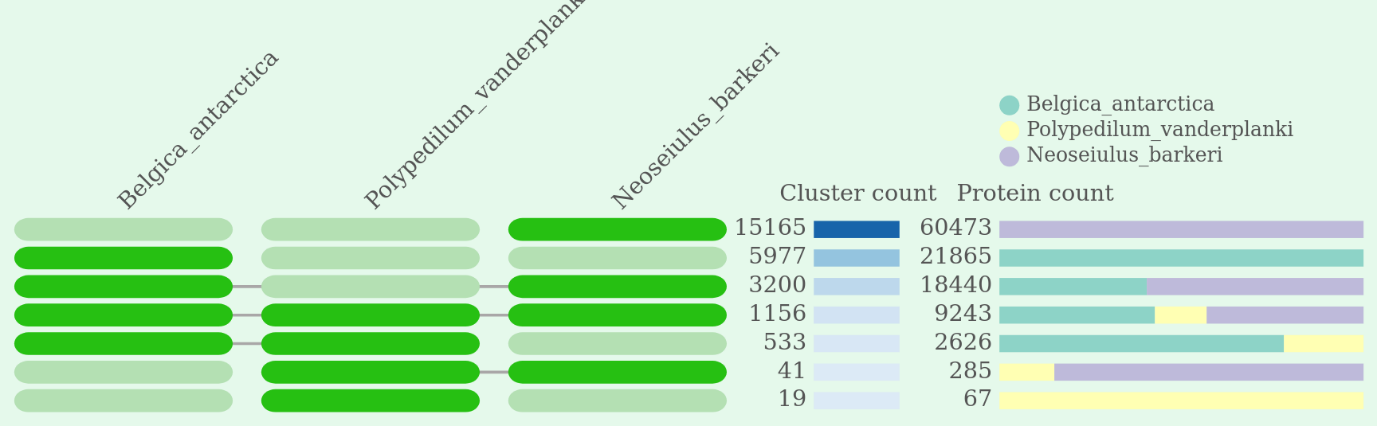


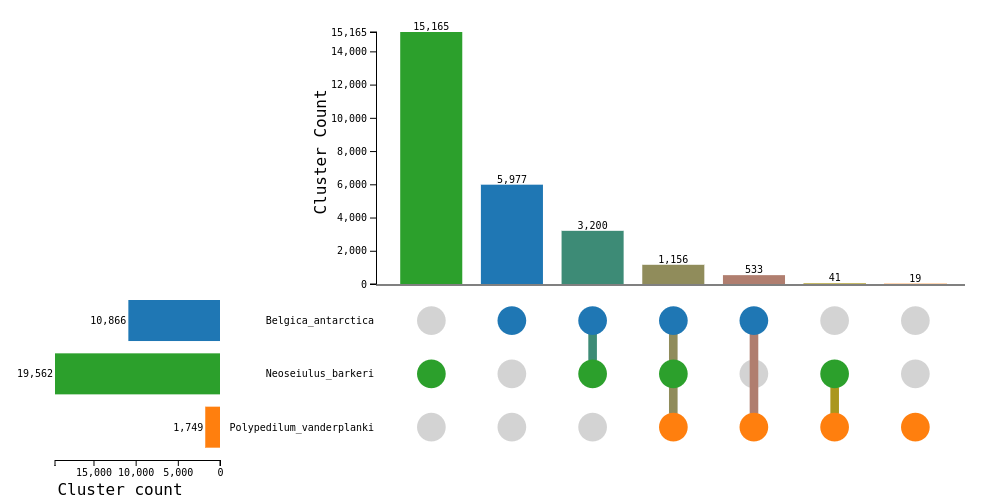




**Results from Orthovenn3 for plants**

**Results from Orthovenn3 for Arachnid**

****

****

The graphical outputs obtained from OrthoVenn3 were instrumental in deciphering meaningful biological insights. The Venn diagrams and interactive cluster visualizations enabled easy identification of core orthologous gene clusters shared across all 21 species, as well as species-specific genes that may reflect unique adaptations or evolutionary events. These visual tools allowed for the detection of gene family expansions and contractions, which are often linked to specialized functions or environmental responses. Functional annotation of these clusters using integrated Gene Ontology (GO) and KEGG pathway enrichment analysis provided deeper insights into the biological roles and molecular functions of conserved gene sets. For instance, clusters unique to specific taxa could hint at lineage-specific traits, while shared clusters might represent fundamental genes essential for basic cellular processes. These comparative visuals not only enhanced our understanding of gene conservation and divergence but also laid the groundwork for exploring functional genomics and evolutionary biology across diverse organisms.

**Article written**

I have written an article about Hi-C sequencing. Hi-C sequencing is a powerful genome-wide technique used to capture the three-dimensional (3D) architecture of chromosomes within the nucleus. In the article I authored, I explored how this method leverages proximity ligation and high-throughput sequencing to identify physical interactions between chromatin regions. By crosslinking DNA, digesting it with restriction enzymes, and re-ligating the resulting fragments, Hi-C generates chimeric reads that reflect spatial proximity rather than linear adjacency. The resulting interaction matrices enable researchers to investigate chromatin loops, topologically associating domains (TADs), and higher-order structural features that are crucial for gene regulation, replication, and genome organization. The article emphasized the growing applications of Hi-C in de novo genome assembly, detection of structural variations, and understanding epigenetic mechanisms, especially in complex plant and animal genomes. Hi-C is proving to be indispensable in both fundamental genome research and applied fields like crop improvement, cancer genomics, and developmental biology.

**Conclusion**

The training experience was an enriching and transformative journey that deepened my expertise in various bioinformatics domains. It provided a multidisciplinary exposure, beginning with molecular docking studies where I explored ligand-receptor interactions using AutoDock4, focusing on the analgesic potential of eugenol in clove oil against dental nociceptors. This hands-on experience in structure-based drug design equipped me with skills in ligand preparation, receptor modeling, and interaction analysis using tools like PLIP, Protein Contact Atlas, and PoseVIEW, enhancing my understanding of computational pharmacology.

In the transcriptomics section, I engaged in differential gene expression analysis using multiple tools including DESeq2, NOISeq, and EdgeR, ultimately applying EdgeR for in-depth statistical analysis in R. This allowed me to identify significantly upregulated and downregulated genes, visualize the data using heatmaps and volcano plots, and interpret the biological relevance of gene expression changes across diverse taxa. The application of R packages sharpened my data handling, visualization, and analytical skills, all essential for transcriptomic research.

My experience extended to orthologous analysis using OrthoVenn3 and OrthoFinder, through which I examined gene conservation and divergence across 21 organisms. The cluster visualizations and comparative genomics insights derived from these tools enabled me to understand gene family evolution, species-specific adaptations, and functional annotations via GO and KEGG pathways. Additionally, constructing phylogenetic trees provided evolutionary context to the genomic data, highlighting relationships among the studied species.

The training also touched upon Hi-C sequencing, where I explored its capability in unraveling the 3D genomic architecture. Understanding its methodology and applications further broadened my perspective on genome organization, chromatin interactions, and structural variation analysis—key elements in modern genomics.

In essence, the diverse set of computational tools and methodologies I practiced during this period not only strengthened my technical proficiency but also empowered me with a holistic perspective of how computational and biological sciences intersect. This training has significantly enhanced my readiness for future research and academic pursuits in bioinformatics, particularly in the fields of comparative genomics, transcriptomics, and computational structural biology.

**References:**

* Cortés-Rojas, D. F., de Souza, C. R. F., & Oliveira, W. P. (2014). Clove (Syzygium aromaticum): a precious spice. Asian Pacific Journal of Tropical Biomedicine, 4(2), 90–96. <https://doi.org/10.1016/S2221-1691(14)60215-X>
* Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: A powerful approach for structure-based drug discovery. Current Computer-Aided Drug Design, 7(2), 146–157. <https://doi.org/10.2174/157340911795677602>
* Berman, H. M., et al. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
* Krivák, R., & Hoksza, D. (2018). P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of Cheminformatics, 10(1), 39. <https://doi.org/10.1186/s13321-018-0285-8>
* O'Boyle, N. M., et al. (2011). Open Babel: An open chemical toolbox. Journal of Cheminformatics, 3, 33. <https://doi.org/10.1186/1758-2946-3-33>
* Morris, G. M., et al. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of Computational Chemistry, 30(16), 2785–2791. <https://doi.org/10.1002/jcc.21256>
* Adasme, M. F., et al. (2021). PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Research, 49(W1), W530–W534. <https://doi.org/10.1093/nar/gkab294>
* Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484
* Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology, 17, 13. <https://doi.org/10.1186/s13059-016-0881-8>
* Xu, L., Dong, Z., Fang, L., Luo, Y., et al. (2019). OrthoVenn3: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research, 47(W1), W52–W58. <https://doi.org/10.1093/nar/gkz333>
* Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology, 20(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>
* Lieberman-Aiden, E., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950), 289–293. https://doi.org/10.1126/science.1181369
* Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics, 14(6), 390–403. https://doi.org/10.1038/nrg3454
* Zhang, X., et al. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nature Plants, 5(8), 833–845. https://doi.org/10.1038/s41477-019-0472-8