UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO
DIPARTIMENTO DI SCIENZE ECOLOGICHE E BIOLOGICHE

Corso di Dottorato di Ricerca in

ECOLOGIA E GESTIONE SOSTENIBILE DELLE RISORSE AMBIENTALI – XXXVI CICLO

Assembly of Robust Transcriptomic Resources for RNA-Seq Data Analysis in Non-Model Organisms: Applications to Ecologically Significant Species

(s.s.d. BIO/11 Biologia Molecolare)

Tesi di dottorato di:
Dott. Pietro Libro

Coordinatore del corso
Prof. Claudio Carere

Tutor
Prof. Daniele Canestrelli

Co-tutor
Prof.ssa Tiziana Castrignanò

Anno 2023/24

Contents

Abstract

This study pioneers the exploration of genetic landscapes in non-model organisms, specifically targeting *Bombina pachypus*, *Hyla sarda*, *Salamandra salamandra*, and *Contracaecum osculatum*, through de novo transcriptome assembly and analysis. By leveraging RNA-Seq data, we aim to unravel the complex genetic underpinnings that govern key ecological and evolutionary traits such as antipredatory behaviors, dispersal mechanisms, and environmental adaptability.

Our research is driven by the need to bridge the knowledge gap in genomic resources for non-model organisms, which play crucial roles in ecological dynamics and biodiversity. The inclusion of *Contracaecum osculatum*, an Antarctic nematode, extends the study to marine parasites, offering a unique perspective on thermal adaptation and survival in extreme environments. This insight is critical for understanding the broader implications of environmental stressors on genetic diversity and organismal resilience.

Through comprehensive transcriptome assembly, the work achieved a detailed view of the genetic blueprint of these species under diverse environmental conditions. The used approach includes rigorous sample collection, RNA preparation, and advanced bioinformatic analyses to ensure the highest quality of data. The resultant transcriptomic resources not only enhance our understanding of specific species but also provide a valuable framework for comparative studies, like with *Pleurodeles watl*, and contribute significantly to the field of ecological genomics.

The findings of this study have profound implications for the understanding of ecological and evolutionary processes. They pave the way for future research in the conservation of biodiversity, the impact of environmental changes on species adaptation, and the exploration of novel genetic pathways in ecological resilience and adaptation strategies. Ultimately, the work underscores the importance of genomic studies in non-model organisms for a comprehensive understanding of life's diversity on Earth.

Chapter 1 - Introduction and objectives

Context

The exploration of genetic landscapes in non-model organisms[1] has become a cornerstone of contemporary biological and ecological research, representing a profound shift in our understanding of life's diversity on Earth. The traditional focus on model species, often more tractable in laboratory settings, has given way to a broader recognition of the ecological significance and evolutionary richness inherent in organisms that do not fit the conventional model organism mold. This paradigm shift is underscored by the advent of cutting-edge technologies, particularly RNA-Seq[2,3] analyses, which empower researchers to delve deep into the transcriptomic intricacies of diverse species.

Evolution of Transcriptomics in Modern Biology

The saga of modern biology unfolds against the backdrop of technological advancements that have revolutionized the study of genes. From the early days of DNA sequencing to the current zenith represented by RNA-Seq technologies, the journey has been marked by an unrelenting quest to decipher the language inscribed in the genetic code. Transcriptomics, as a discipline, has transcended the constraints of model-centric approaches, heralding a new era where the exploration of genetic landscapes extends to the vast realm of non-model organisms. This transformative trajectory is not merely a chronological progression; it signifies a profound shift in our philosophical and methodological approach to biological exploration.

The transcriptome, encompassing the entirety of RNA molecules present in a cell at a given moment, stands as a testament to this evolution. Composing mRNAs carrying genetic information for protein coding, along with various other intracellular RNA forms with predominantly regulatory functions, the transcriptome reflects the unique characteristics of a single cell[4] in a specific temporal or environmental condition. Transcriptomic analysis delves into the entire expression profile of a biological sample, quantitatively and qualitatively characterizing molecular events with clinical impact in coding regions[5]. Additionally, it identifies changes in transcript processing or stability, shedding light on structural genome rearrangements at the messenger RNA level.

One of the most intriguing applications of quantitative transcriptomics is the analysis of differential gene expression, aiming to identify genes exhibiting significant differences in

expression levels among experimental conditions[6,7,8]. The statistical verification of differential expression involves hypothesis tests, evaluating whether the observed difference between two sample groups can be considered statistically significant[9,10]. This method finds effective application in the study of conditions such as muscle diseases and syndromic conditions associated with multisystemic clinical frameworks. In both healthy and diseased subjects, transcriptomic information allows the detection of genes expressed significantly differently between the two groups, highlighting differences induced by pathological conditions[11].

The realm of pharmacology, especially in cancer treatment, has witnessed substantial progress through transcriptomic studies. Personalized therapies targeting specific and individual mutations in various cancer types have become a reality, showcasing the potential of transcriptomic analysis in tailoring treatments to the unique genetic makeup of individuals. Additionally, transcriptomic analysis aims to monitor spatiotemporal variations in the expression of all genes in an individual, offering insights into responses to stress and identifying genes playing a primary role in stress tolerance/susceptibility.

Beyond the clinical domain, transcriptomics has played a pivotal role in advancing ecological understanding[12,13]. In the study of vertebrates like amphibians, it has contributed significantly to unravelling morphological and ecological adaptations that form the basis for their occupation of diverse terrestrial environments[14,15].

This narrative encapsulates the essence of a paradigm shift in biology, where technological innovations, particularly in RNA-Seq, not only provide intricate datasets but also redefine the boundaries of biological exploration.

Crucial Role of RNA-Seq Analyses

In recent years, the scientific landscape of genetics has witnessed a transformative wave with the advent and rapid development of Next Generation Sequencing (NGS) platforms[16], heralding a new era in genomic exploration. These platforms, capable of parallel DNA sequencing, have not only significantly reduced the costs but also the turnaround times for sequencing individual genes, gene panels, exomes, and entire genomes. The implications of these advancements are profound, promising more in-depth and efficient molecular investigations.

For decades, DNA sequencing relied primarily on Sanger's methodology[17], developed in the 1950s, which predominantly analyzed single-stranded DNA. Unlike NGS, Sanger sequencing lacks the ability to execute parallel sequencing. Consequently, when dealing with the sequencing of extensive genomic regions or a high number of genes, Sanger sequencing incurs considerably greater time and cost.

RNA-Seq, an abbreviation for RNA sequencing, represents a revolutionary next-generation sequencing technique that utilizes NGS to unveil the presence and quantity of RNA in a biological sample at a given moment. This method scrutinizes the ever-evolving cellular transcriptome. RNA-Seq's capabilities extend to examining alternative splicing transcripts, post-transcriptional modifications, gene fusions, mutations/SNPs, and changes in gene expression over time or differences in gene expression among various groups or treatments. Furthermore, RNA-Seq can explore various RNA populations, including total RNA, small RNA (e.g., miRNA), tRNA, and ribosomal profiling.

The power of RNA-Seq is evident in its ability to determine exon/intron boundaries and verify or amend previous annotations of 5' and 3' regions in genes. Recent advancements in RNA-Seq encompass single-cell sequencing, in situ sequencing of specific tissues, and sequencing native RNA molecules in real-time at the single-molecule level. Emerging applications driven by progress in bioinformatic algorithms include copy number alterations, microbial contamination, transposable elements, cell type deconvolution, and the presence of neo-antigens.

RNA-Seq emerges as a potent method for discovering, profiling, and quantifying RNA transcripts. In an RNA-Seq experiment, reads constitute raw data from which information about gene expression levels in the sample is derived. These reads, originating from the reverse transcription of mRNA into cDNA and subsequently amplified through cloning techniques, serve as the foundational data for gene expression analysis.

To quantify the number of reads attributed to each gene, the reads from each sample are mapped to a reference genome or transcriptome[18]. The alignment of reads to the reference, performed by software, involves finding the position in the genome or transcriptome where the read obtains the best match or the highest sequence homology. Once the positions of reads on the reference are determined, it is possible to count the number of reads aligned to a

gene (or transcript or exon). The total reads aligned to a gene are referred to as counts, representing a measure of the gene's expression level.

To obtain a precise estimate of expression levels for different genes in a biological sample, the count of the number of reads must be normalized, considering the length of the gene under examination and the total number of reads mapped to the sample[19,20]. The counts are the final data of an RNA-Seq experiment for transcriptome quantification and are stored in matrices where rows represent different genes, and columns represent different sequenced samples.

Statistically, counts, by definition, represent a sum of independent random events (the mapping of each read to genes). They can, therefore, be described by a random variable following a specific statistical distribution. The two most used distribution models to describe counts are the Poisson model and the Negative Binomial model[21]. However, it has been demonstrated that the stringent condition of equality between mean and variance, as posited by the Poisson model, is insufficient to adequately describe the variability of RNA-Seq data. If the variance exceeds the mean, the data is considered over dispersed, rendering the Poisoning distribution assumption inappropriate[22]. To address this, a count model based on a negative binomial distribution was proposed, introducing the φ factor representing the coefficient of variation of biological variability across different subjects. This factor allows the separation of the effects of biological variability from those due to technical variability.

These advancements in sequencing technologies, especially RNA-Seq, have opened unprecedented avenues for molecular investigations, providing a wealth of complex and intricate data. Managing this information requires dedicated procedures for data treatment and preservation, including specialized databases, to be employed with clear policies safeguarding the interests of all users. The integration of these technologies into molecular studies offers a transformative potential for scientific exploration, significantly enhancing our ability to dissect and comprehend the intricate molecular landscapes that govern life's diversity.

Advantages of Transcriptomics Research

Transcriptomics, as a pivotal realm within the vast landscape of molecular biology, offers an array of distinctive advantages that propel scientific inquiry to unprecedented heights. One of

its central merits lies in the nuanced exploration of gene expression patterns across diverse biological contexts. This dynamic approach enables researchers to unravel the intricate tapestry of cellular responses, capturing the essence of how organisms adapt and evolve in response to internal and external stimuli.

Furthermore, transcriptomics serves as a potent tool for deciphering the intricacies of alternative splicing, post-transcriptional modifications, and the diverse array of RNA species beyond messenger RNA (mRNA). By delving into the full spectrum of RNA molecules, including transfer RNA (tRNA), ribosomal RNA (rRNA), micro-RNA (miRNA), and long non-coding RNA, transcriptomics provides a comprehensive understanding of the regulatory networks orchestrating cellular functions.

The advent of high-throughput sequencing technologies, epitomized by Next Generation Sequencing (NGS) platforms, has revolutionized transcriptomics research. These platforms empower scientists to conduct large-scale analyses with unprecedented speed and cost-efficiency. The seamless parallel sequencing capabilities of NGS platforms have not only accelerated data acquisition but have also democratized access to transcriptomic investigations, extending the scope of research beyond traditional model organisms to encompass a broader array of species.

Moreover, transcriptomics unfolds its prowess in unrevealing the subtleties of gene regulation and expression dynamics. The exploration of non-coding regions, often overlooked in traditional genetic studies, becomes a focal point, unveiling potential regulatory elements and novel functional insights. This holistic approach to gene expression profiling provides a panoramic view of the molecular landscape, offering a richer understanding of the underlying molecular mechanisms steering biological phenomena.

In essence, transcriptomics emerges as a beacon guiding researchers through the intricate complexities of genetic information. Its advantages extend beyond the traditional boundaries of molecular research, fostering a profound comprehension of the molecular symphony orchestrating life's myriad processes. As the technological tapestry of transcriptomics continues to evolve, its contributions to biology promise to be both enduring and transformative.

Challenges in the Study of Non-Model Organisms

A mere decade ago, transcriptomes were exclusively sequenced for organisms deemed of paramount interest in scientific research, the so-called model organisms. However, the

technological strides made in 2010 with high-throughput sequencing, also known as next-generation sequencing, have not only proven cost-effective but have also streamlined the labor-intensive aspects of sequencing. Consequently, the repertoire of organisms investigated using these methods is steadily expanding.

The examination of non-model organisms holds the promise of unveiling fresh insights into the mechanisms underpinning the "diversity of fascinating morphological innovations" that have fostered life's abundance on Earth. In the realms of animals and plants, these innovations, which include mimicry, mutualism, parasitism, and asexual reproduction, often elude scrutiny in common model organisms. De novo transcriptome assembly has emerged as the preferred method for studying these non-model organisms[23]. It is an economical and simpler alternative to genome construction, especially since model organism analysis methods are inapplicable without an existing genome.

Transcriptomes of these non-model organisms have the potential to reveal new proteins and their isoforms implicated in these unique biological phenomena. Despite the scarcity of molecular studies on non-model organisms compared to their model counterparts, these non-model organisms constitute most species in nature.

The potency of such studies is underscored by the increasing number of RNA-seq samples deposited in the NCBI Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) database over recent years. Focusing on the taxa "amphibia", the observation of a marked increase in RNA-seq samples over a mere two months is noteworthy, particularly when contrasted with the relatively stagnant numbers of organisms with sequenced genomes or de novo assembled transcriptomes. This data exemplifies the vast potential of transcriptomic studies, particularly in the exploration of differentially expressed genes, in non-model organisms where a reference genome is absent.

Considering the challenges posed by limited genomic resources in non-model organisms, this study takes a pioneering step by spearheading the assembly of robust transcriptomic resources. This strategic initiative not only addresses these challenges but also lays the groundwork for more comprehensive ecological and evolutionary investigations.

Ecological Relevance of the Selected Species

The meticulously chosen species - *Bombina pachypus*, *Hyla sarda*, *Salamandra salamandra*, and *Contracaecum osculatum* - serve as emblematic representatives of diverse ecological

niches, each contributing a unique dimension to our exploration of ecological dynamics and biodiversity patterns.

Bombina pachypus

The Apennine yellow-bellied toad, *Bombina pachypus*, stands as a vibrant and endemic amphibian in the Apennine Peninsula, offering a captivating intersection of ecological significance and evolutionary charm. Beyond being a subject of study, this toad becomes a representative of ecological resilience and adaptability, thriving in diverse habitats ranging from forests to freshwater systems[24].

This unique amphibian draws attention with its distinctive behavioural and physiological traits, making it a focal point for research into the evolutionary dynamics of defensive strategies and environmental adaptation. Notably, *Bombina pachypus* is renowned for its 'unken reflex' – a defensive behaviour where it exposes its brightly coloured belly as a warning to potential predators[25-28]. This, coupled with its aposematic coloration, provides an intriguing model for exploring the evolutionary origins and genetic foundations of antipredator strategies.

In the broader context of ecological and evolutionary research, *Bombina pachypus* serves as a witness to the adaptive complexity of amphibians. Its life cycle, encompassing both aquatic larval stages and terrestrial adulthood, presents an exceptional opportunity to study the genetic and physiological adaptations required to thrive in such varied environments.

As an indicator species, the health and behaviour of *Bombina pachypus* reflect broader environmental conditions, making it a crucial subject for ecological monitoring and conservation efforts[29]. The primary goal of this research concerning *Bombina pachypus* is to develop a comprehensive understanding of its genetic composition and expression patterns.

Utilizing advanced RNA sequencing techniques and de novo transcriptome assembly, we aim to unveil the genetic foundations of its unique behavioural traits and adaptive mechanisms. This endeavour goes beyond gene cataloguing; it involves piecing together the evolutionary story of a species that has much to teach us about survival, adaptation, and the intricate interplay between genetics and ecology.

Hyla sarda

Inhabiting the arboreal landscapes of the Mediterranean islands, the Tyrrhenian tree frog, *Hyla sarda*, emerges as a pivotal subject in ecological and evolutionary biology[30]. Endemic

to the islands of Sardinia and Corsica, this species presents an intriguing study in understanding island biogeography, arboreal adaptation, and acoustic communication in amphibians.

The adaptation of Hyla sarda to arboreal habitats stands as a testament to its evolutionary ingenuity. Residing primarily in tree canopies, this frog has developed unique morphological and behavioural traits suited for life above the ground. Its slender body, long limbs, and distinctive coloration not only facilitate its arboreal lifestyle but also contribute to its role in the ecosystem as both predator and prey. This vertical niche occupation offers a unique perspective on how organisms adapt to specialized habitats, both morphologically and behaviourally[31-34].

The vocalizations of *Hyla sarda*, particularly prominent during the breeding season, serve not only as a means of communication but also as a critical element in the study of animal behaviour and mating strategies. The male mating call, a crucial component for reproductive success, provides an excellent model for studying the genetic basis of bioacoustics communication and mate selection in amphibians. Understanding the genetic foundations of these vocalizations can offer insights into the evolution of communication strategies in response to environmental pressures and sexual selection.

Our research objectives regarding *Hyla sarda* are multifaceted. Firstly, we aim to decode the genetic architecture supporting its unique arboreal adaptations and acoustic communication strategies. Through advanced genomic sequencing techniques and bioinformatic analysis, we seek to uncover the genetic blueprint orchestrating these specialized traits. Additionally, given its distribution in isolated island ecosystems, *Hyla sarda* provides an unprecedented opportunity to study the effects of geographic isolation on genetic diversity and speciation.

Furthermore, this research holds critical importance for conservation purposes. Faced with threats from habitat loss, pollution, and climate change, understanding the species' genetic adaptability and resilience is vital for devising effective conservation strategies. By exploring genetic diversity within and among populations, we can gather information about the species' ability to respond to environmental changes, aiding in the development of tailored conservation plans for its unique ecological needs.

Salamandra salamandra

The Fire Salamander, *Salamandra salamandra*, takes center stage in the realm of European herpetofauna, renowned for its striking coloration and remarkable biological features. This

species, adorned with vivid black and yellow markings, not only captivates the imagination but also serves as a critical subject in our exploration of ecological and evolutionary processes.

The habitat of the Fire Salamander spans various European landscapes, from the Iberian woodlands to the lush forests of the Balkans. This broad geographic distribution, coupled with its distinct ecological adaptations, positions it as an ideal model for studying genetic diversity and adaptive evolution[35-38]. Particularly noteworthy is its unique reproductive strategy: viviparity, wherein it gives birth to live larvae, a rarity among amphibians. This aspect alone opens a window into understanding the intricate genetics of reproduction and development.

Beyond its reproductive uniqueness, the Fire Salamander is also renowned for its longevity and its ability to produce a variety of toxic compounds as a defense mechanism[39,40]. These traits offer unprecedented opportunities to delve into the genetic mechanisms underlying toxin production and longevity, contributing to our broader understanding of these biological phenomena.

The primary goal of our research concerning the Fire Salamander is to elucidate the genetic foundations underlying its ecological and physiological adaptations[41,42]. Through cutting-edge genomic sequencing and analysis, we aim to decode the genetic blueprint governing its unique characteristics. This investigation is not merely an academic pursuit; it holds profound implications for conservation biology. Understanding the genetic basis of the Fire Salamander's adaptability and resilience is crucial for developing strategies to safeguard it from growing threats such as habitat loss, climate change, and epidemics.

Furthermore, this research extends beyond individual species to contribute to our understanding of amphibian biology and ecology, offering insights that transcend the boundaries of a single species and enrich our broader comprehension of these fascinating creatures.

Contracaecum osculatum sp. D

Embarking into marine ecosystems, *Contracaecum osculatum sp. D*, an Antarctic nematode, unfolds as a captivating subject within parasitology and evolutionary biology. Predominantly found in marine mammals' and fish digestive tracts, this organism yields valuable insights into host-parasite interactions, adaptation to extreme environments, and the evolutionary pressures shaping parasite life cycles[43-45].

The life cycle of *Contracaecum osculatum* sp. D is intricate, involving multiple hosts, showcasing sophisticated evolutionary adaptation to its parasitic lifestyle. Its ability to thrive in diverse hosts underscores the genetic and physiological versatility intrinsic to successful parasites, significantly influencing our understanding of parasite-host dynamics and the impact of parasites on marine ecosystems[46].

A key objective in researching *Contracaecum oscul*atum sp. D is unraveling the genetic foundations behind its adaptability and survival mechanisms in various marine environments. Leveraging advanced genomic sequencing and analysis, our aim is to dissect the genetic architecture enabling this nematode to navigate its parasitic life cycle successfully[47,48]. This exploration promises in-depth insights into the crucial genes and molecular pathways governing parasites' adaptation and survival. Another critical facet of studying *Contracaecum osculatum* is its role in marine food webs and potential implications for biodiversity and marine health. As a parasite, it not only influences the health and behavior of its hosts but also serves as an indicator of ecosystem health.

This research broadens the study's horizons, accentuating the unique contributions of marine parasites to genetic diversity and adaptation across varied habitats. With each species playing a distinctive ecological role, our exploration continues to deepen our comprehension of the intricate interplay between genes and environments, expanding our knowledge of the diverse strategies organisms employ to thrive in challenging ecological niches.

Research objectives

In this comprehensive study, our overarching research objectives span across emblematic amphibian and nematode species, delving into their genetic landscapes to unravel intricate ecological and evolutionary dynamics. Our exploration commences with the Apennine toad, *Bombina pachypus*, a charismatic endemic amphibian that epitomizes a fascinating intersection of ecological relevance and evolutionary charm. Focusing on the genetic underpinnings of its unique behavioral and physiological traits, we aim to not only catalog its genes but also weave a narrative of its evolutionary history. Through advanced RNA sequencing and de novo transcriptome assembly, our goal is to decode the genetic intricacies that underscore its defensive strategies, adaptive capabilities, and resilience.

Transitioning to the Tyrrhenian tree frog, *Hyla sarda*, our research amplifies its significance as an invaluable model for island biogeography, arboreal adaptation, and bioacoustics communication. By deciphering the genetic foundations orchestrating its arboreal lifestyle

and distinctive vocalizations, we aim to glean insights into the adaptive pathways that have shaped its unique traits. Utilizing cutting-edge transcriptome sequencing and bioinformatic analyses, our objectives extend beyond academic inquiry; we aspire to contribute to conservation efforts by comprehending its genetic adaptability in the face of environmental challenges.

The Fire Salamander, *Salamandra salamandra*, emerges as a pivotal subject in our exploration, showcasing the key to unraveling genetic diversity and adaptive evolution. Beyond its aesthetic allure, this species provides a window into the complex genetic underpinnings of its ecological and physiological adaptations. Employing state-of-the-art transcriptome sequencing and analysis, our research strives to unveil the genetic architecture governing its distinctive traits. However, our ambitions transcend academic curiosity; we endeavour to apply our findings to inform conservation strategies tailored to safeguarding this species from looming threats.

Venturing into marine ecosystems, *Contracaecum osculatum*, an Antarctic nematode, widens the scope of our study to include marine parasites. Our primary objective is to unravel the genetic bases of its adaptability and survival in diverse marine environments. Through advanced transcriptome sequencing and analysis, we seek to dissect the genetic intricacies that enable this nematode to navigate its complex life cycle successfully. Simultaneously, we explore its role in marine food webs, unreaveling potential implications for biodiversity and marine health.

Collectively, our research objectives transcend the mere cataloguing of genes; they encapsulate a holistic understanding of the adaptive strategies, evolutionary trajectories, and ecological roles these species embody. By navigating genetic frontiers across diverse ecological niches, our study aims to contribute profoundly to the broader fields of ecology, evolution, and conservation, enriching our comprehension of the intricate interplay between genes and environments.

Methodological Approach

The methodological approach followed is meticulously designed to address the complexities and nuances of conducting transcriptomic analysis in non-model organisms. The foundation of this approach is predicated on a comprehensive research design that meticulously intertwines various scientific techniques and theories to unravel the intricacies of genetic expression in these organisms.

The journey of this research begins with a carefully orchestrated process of selecting the non-model organisms. As previously mentioned, this selection was not arbitrary; instead, it was guided by a set of stringent criteria that ensured the organisms were not only representative of the study's objectives but also feasible for the required experimental procedures. The process of sample collection was equally rigorous, involving precise RNA preparation techniques that adhered to the highest standards of scientific excellence. This phase was critical, as the integrity of the RNA samples would later play a pivotal role in the reliability of our findings.

Once the samples were secured, the focus shifted to the RNA-Seq data analysis pipeline. This stage was the crux of our methodological approach, involving a series of complex steps to process the RNA-Seq data. Quality control was paramount, ensuring that the raw data was pristine before it underwent the de novo transcriptome assembly process. Here, we employed a suite of sophisticated software and algorithms, detailed in Chapter 3, each chosen for their proven efficacy in handling such intricate data. The validation of the assembled data was not an afterthought but a key step in our methodology, ensuring that our results were not just accurate but also reproducible.

The narrative of our methodology further extends to the realm of bioinformatics tools and statistical analysis. In this domain, we utilized an array of advanced tools specifically tailored for transcriptome analysis. These tools were instrumental in identifying differentially expressed genes, annotating them, and performing a detailed functional and pathway analysis. Our choice of statistical methods was equally deliberate, designed to provide a robust framework for data interpretation, thereby reinforcing the credibility of our findings.

However, our methodological approach was not devoid of challenges and limitations. These hurdles, inherent in any scientific endeavor, were acknowledged and addressed throughout the study. This candid appraisal of our methodology's constraints not only underscores the authenticity of our research but also opens avenues for future improvements.

Quality control was not just a procedural formality but a fundamental aspect of our approach, ensuring that every facet of our study, from sample collection to data analysis, was executed with precision and integrity.

In summary, the methodological approach of this doctoral thesis represents a harmonious blend of scientific rigor and a deep understanding of the challenges associated with transcriptomic analysis in non-model organisms. It is a testament to the meticulous planning and execution that underpins this research endeavor.

Chapter 2 - Theory and Best Practices for transcriptomic analysis of non-model organisms

Introduction

The advancement of RNA-Seq technologies is closely intertwined with the development of sophisticated bioinformatics tools. The vast amount of data generated by RNA-Seq requires powerful computational methods for analysis and interpretation. Bioinformatics tools have been developed and refined to manage, analyze, and visualize RNA-Seq data, enabling researchers to extract meaningful biological insights. These tools have facilitated the identification of gene expression signatures, the discovery of novel genes and isoforms, and the understanding of complex regulatory networks. While bulk RNA-Seq offers a powerful means to study gene expression, it also comes with certain considerations. The technique excels in situations where the goal is to decipher the dominant gene expression trends in a population of cells, providing insights that are representative of the whole. However, it may mask the heterogeneity and distinct expression profiles of individual cells within the sample. Despite this limitation, bulk RNA-Seq remains a widely used and efficient method, especially in contexts where single-cell resolution is not critical. It continues to be a valuable tool in the arsenal of biologists, offering a cost-effective and robust way to explore the complexities of gene expression in various biological systems. The interpretation of RNA sequencing (RNA-Seq) data presents a significant challenge, primarily due to its complexity and the sheer volume of information it generates. Each RNA-Seq experiment produces millions of short reads, representing fragments of the transcriptome from a biological sample. This high-throughput nature results in a vast dataset that requires meticulous analysis to decipher. The challenge is not only in the quantity of data but also in its complexity. RNA-Seq data encompasses a wide array of RNA types, including mRNAs, non-coding RNAs, and a multitude of splice variants, each contributing to an intricate picture of the transcriptome. This diversity necessitates sophisticated computational tools and algorithms to accurately assemble, align, and quantify these sequences to draw meaningful biological conclusions.

Another layer of complexity in RNA-Seq data interpretation arises from the need for advanced bioinformatics expertise. The initial steps of quality control, sequence alignment, and assembly are just the beginning. Researchers must navigate through various bioinformatic hurdles, including differentiating between biologically significant signals and technical noise, identifying novel transcripts, and accurately quantifying gene expression

levels. Furthermore, the biological interpretation of RNA-Seq data is non-trivial. It requires a deep understanding of the biological context to translate these data into insights about gene function, regulation, and interaction. The dynamic nature of the transcriptome, influenced by factors such as developmental stage, environmental conditions, and disease states, adds another layer of complexity to the analysis and interpretation of RNA-Seq data.

De novo transcriptome assembly is a critical process in RNA sequencing studies, especially when a reference genome is unavailable or incomplete. This method involves piecing together RNA-Seq reads to reconstruct the full-length transcripts that represent the expressed genes in a sample. The process starts with the generation of millions of short sequences reads, which are then assembled into longer sequences called contigs. The key challenge in de novo assembly is to merge these short reads accurately and efficiently without the guidance of a reference genome, relying solely on the overlap of sequences to infer the original transcript structure. This approach is invaluable for studying non-model organisms or exploring genomic regions that are poorly characterized, allowing researchers to uncover novel genes and transcripts and understand their functional implications.

The complexity of de novo transcriptome assembly lies in its reliance on sophisticated computational algorithms and tools. Assemblers use a variety of techniques, most notably the De Bruijn graph method[49], which constructs a network of sequence overlaps represented as a graph. Key tools in this space, such as Trinity[50], SOAPdenovo-Trans[51], and Oases[52], utilize this approach, enabling them to handle the vast amount of data and the intricacies of transcriptome assembly. The choice of k-mer length, which represents the size of sequence overlaps used to build the graph, is a critical decision that influences the accuracy of the assembly[53]. A longer k-mer length may lead to more specific but potentially incomplete assemblies, while shorter k-mers may increase the risk of erroneous assembly due to more frequent overlaps. Assemblers must strike a balance, and often the choice of k-mer length is optimized based on the nature of the sample and the sequencing data.

Following the intricate process of de novo transcriptome assembly, it's crucial to engage in rigorous post-assembly quality control and refinement. This stage is pivotal to ensure the integrity and usability of the assembled transcriptome for further scientific inquiry. The immediate task is to assess the quality of the assembly, which involves evaluating metrics like contig length distribution and coverage depth. Tools like BUSCO[54] (Benchmarking Universal Single-Copy Orthologs) play a vital role here, as they provide insights into the

completeness of the assembly by comparing it against a set of evolutionarily conserved genes. This assessment helps in identifying potential gaps, mis assemblies, or chimeric sequences. Further refinement might involve filtering out low-quality or redundant sequences and possibly reassembling fragments to produce a more coherent and comprehensive transcriptome. These steps are fundamental in ensuring that the final assembled transcriptome is a true and effective representation of the RNA present in the original sample.

Annotation is the next critical step in the post-assembly process, transforming the raw sequence data into a functionally informative resource. This phase involves assigning biological meaning to the sequences, such as identifying coding regions, predicting gene functions, and classifying non-coding RNAs. Annotation can be complex, as it often requires the integration of various computational tools and databases that compare the newly assembled sequences to known genes, proteins, and functional motifs. Techniques like homology search and domain prediction are commonly employed, linking sequences to known biological functions and pathways. This annotated transcriptome becomes a valuable resource for various downstream analyses, including differential gene expression studies, discovery of novel genes or isoforms, and comparative genomic studies. The quality and depth of annotation directly impact the biological insights that can be derived from the transcriptome, making it a crucial step in the journey from sequencing reads to meaningful biological understanding.

Following images and captions in this chapter are adapted from “A Simple Guide to De Novo Transcriptome Assembly and Annotation. Academic Press”[55].

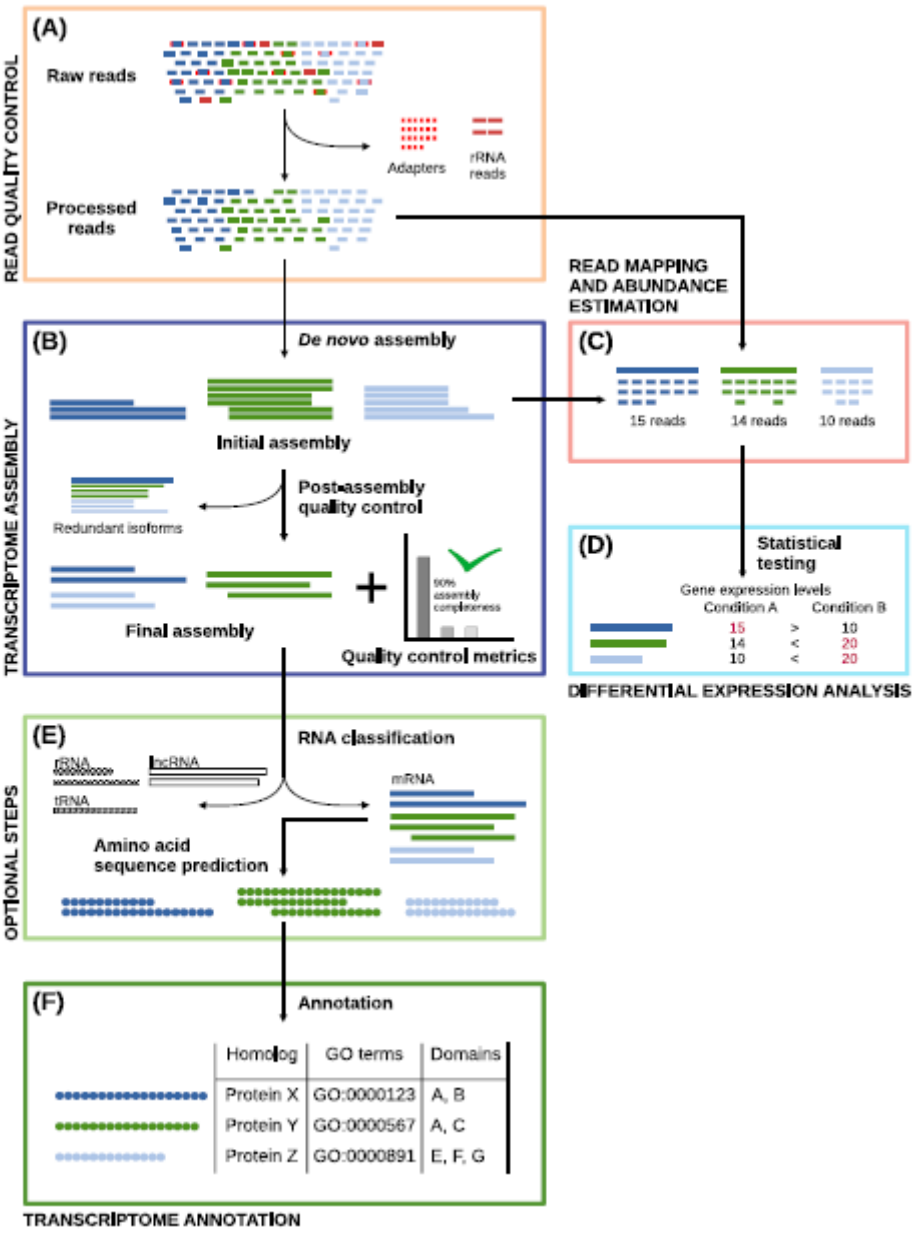


Figure 1 –Workflow for Assembly and Annotation. (A) Initial quality control involves filtering out erroneous reads and sequencing artifacts from the raw data. (B) The process of sequence assembly encompasses the clustering of isoforms (transcript variants that result from alternative splicing) and the elimination of superfluous sequences. (C) The assembled sequences undergo mapping with the original raw reads, serving a dual purpose: to verify the quality of the assembly and to facilitate differential expression analysis. (D) Statistical methods are employed to detect changes in expression levels. (E) The sequences are categorized based on RNA types and translated into protein sequences prior to annotation. (F) The annotation phase involves identifying sequences by their similarity, pinpointing sequence features like functional domains, and assigning Gene Ontology terms.

Pre-assembly quality control and filtering

Before embarking on the complex task of de novo transcriptome assembly, it is essential to ensure that the input RNA-Seq data is of the highest quality. This initial phase, known as pre-assembly quality control, involves a thorough examination and refinement of the raw sequencing reads. These reads, although generated through advanced sequencing technologies, are not free from errors and inconsistencies. Factors such as sequencing errors, variations in read length, and the presence of non-biological sequences (like adapters) can significantly impact the accuracy of the assembly. Therefore, rigorous quality control is crucial to filter out these inaccuracies, setting a solid foundation for a reliable assembly process.

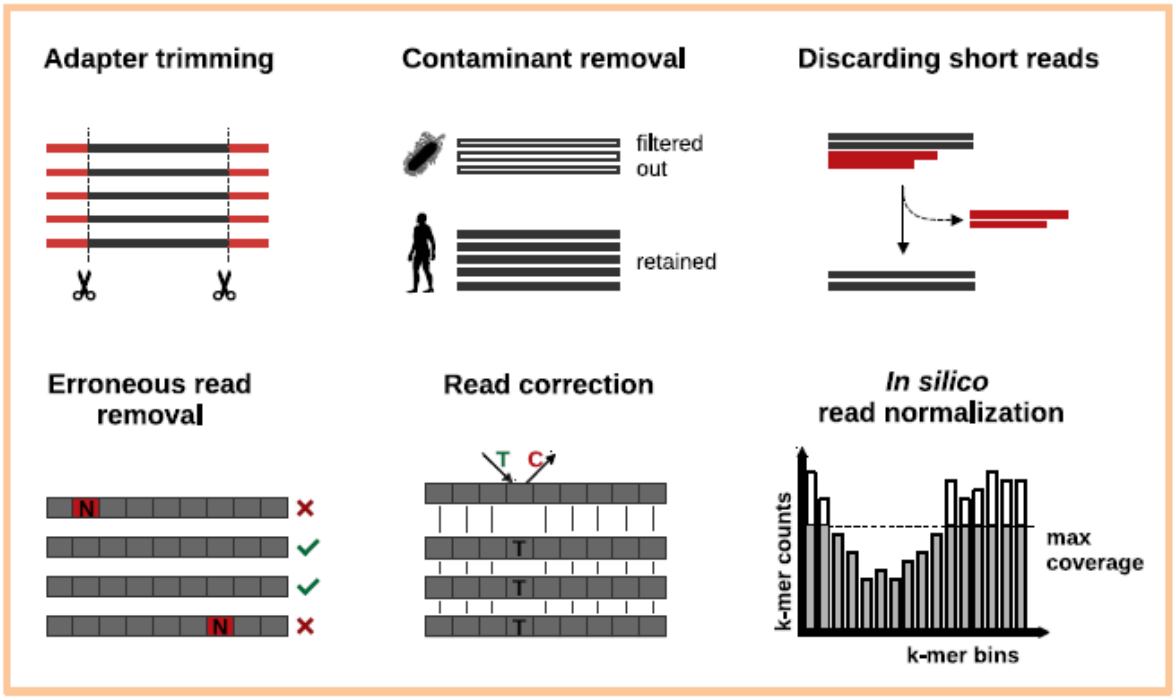


Figure 2 - Illustrates the process of quality control and data cleansing for short reads. This involves several steps, such as trimming adapters, discarding short and erroneous reads that contain N-bases, and correcting reads through comparison with other sequences. Additionally, it includes the exclusion of reads that stem from contaminating sources, for example, pathogens in host species. For extremely large datasets (exceeding 200 million reads), in silico read normalization is a beneficial preprocessing step. It can markedly enhance the performance of the assembler by selectively downsizing the number of reads while preserving the original dataset's transcriptomic complexity.

The first step in pre-assembly quality control is the comprehensive inspection of sequencing reads. Tools like FastQC[56] and Falco[57] provide a detailed overview of various quality metrics, such as per-base sequence quality, GC content, sequence length distribution, and the presence of overrepresented sequences. This analysis helps in identifying potential issues like low-quality bases, adapter contamination, and sequence duplication. A careful examination of these reports informs subsequent steps of read cleaning and filtering, ensuring that only high-quality reads are used for assembly.

Following the initial assessment, the next critical step is read cleaning. This involves trimming or removing low-quality bases and sequences, which can skew the assembly process. Adapter sequences, artificially added during library preparation for sequencing, must be meticulously removed as their presence can lead to erroneous assembly. These kinds of tools are adept at detecting and excising these adapter sequences, while also offering functionalities like length-based sequence filtering, which is crucial for discarding extremely short reads that may result from this process.

An often-overlooked aspect of pre-assembly quality control is the identification and removal of contaminants. These could be sequences from foreign sources (like bacterial or fungal contaminants in a eukaryotic sample) or cognate contaminants like rRNA sequences. Additionally, in datasets with a high number of reads, in silico normalization can be performed. This process reduces the redundancy of reads representing well-covered transcripts while retaining the complexity of the transcriptome, thereby optimizing the subsequent assembly process.

The pre-assembly quality control and filtering stage is more than just a preparatory step; it's a critical determinant of the success of the entire transcriptome assembly process. By rigorously cleaning and preparing the sequencing data, researchers can significantly reduce the computational complexity and enhance the accuracy of the assembly. This phase, although time-consuming and demanding in terms of computational resources, is indispensable for ensuring that the assembly process is based on the most accurate and representative dataset possible.

De novo transcriptome assembly

De novo transcriptome assembly is a pivotal process in genomics, particularly in the study of organisms for which no reference genome is available. This method involves the

reconstruction of transcripts directly from short RNA-Seq reads, without relying on a pre-existing genome as a guide. The primary objective of de novo assembly is to accurately and comprehensively piece together these short sequences into longer contiguous sequences, or contigs, which represent the transcriptome of the organism. This process is essential for exploring genetic diversity, discovering novel genes, and understanding gene expression in organisms not previously well-characterized at the genomic level.

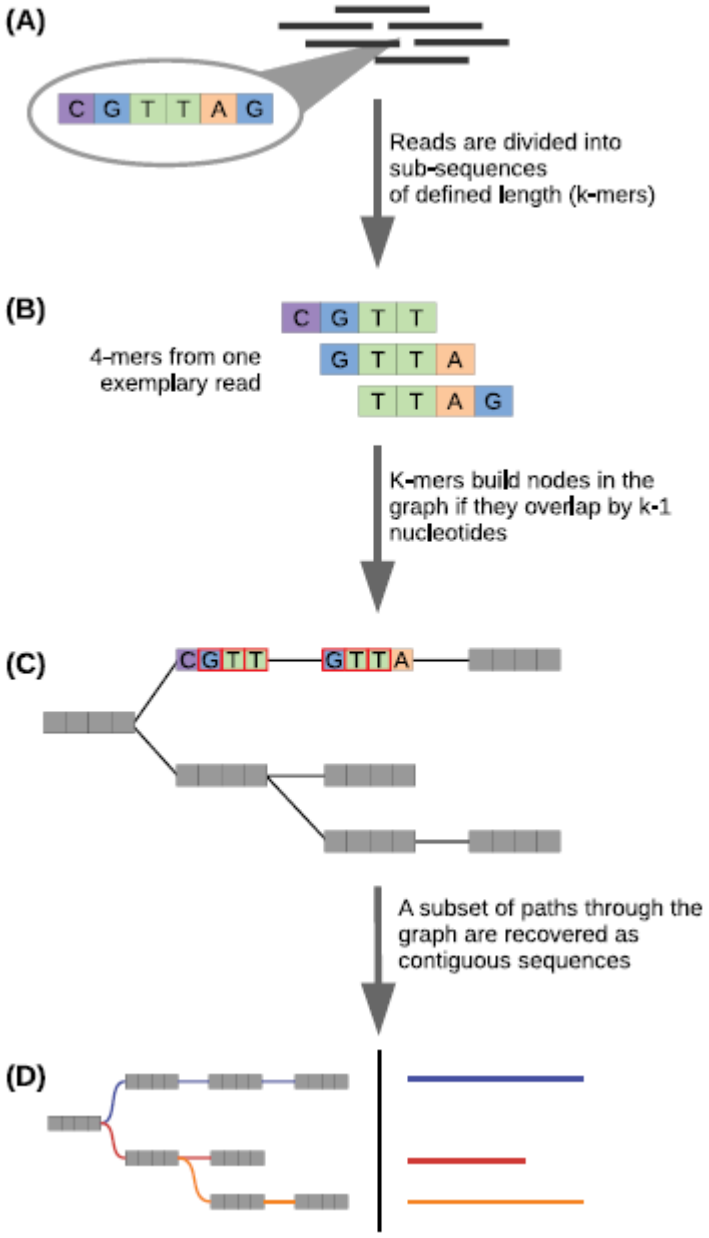


Figure 3 – A Graph-Based Method for De Novo Transcriptome Assembly. This method involves creating a collection of sub-strings from RNA-Seq reads and arranging them into a graph or a series of graphs. In these graphs, sub-strings are linked if they overlap, thus forming pathways that represent the original transcripts from which the reads may have been derived. (A) The assembly process starts with short nucleotide reads

ranging from 50 to 250 nucleotides. In the case of paired-end reads, the pairs are combined into a single, continuous read before assembly. For instance, a segment of 6 nucleotides (CGTTAG) from a read is highlighted here. (B) From each read, all possible subsequences of a certain length k (known as k-mers) are generated. For example, the 4-mers (when k equals 4) derived from the previously mentioned 6-nucleotide fragment are shown. (C) Each k-mer then forms a node (or vertex) in the graph. Edges, or connections, are created between any two nodes that overlap by k-1 nucleotides. This step is crucial for establishing the graph's structure, as illustrated by the edge linking the first two 4-mers from the earlier step in a simplified De Bruijn graph. (D) The final stage involves traversing various paths through the graph(s) and identifying them as distinct sequences. However, not all paths are selected; only those representing valid transcripts are algorithmically chosen.

The assembly process begins with an enormous volume of short reads, each a fragment of the organism's RNA. The main challenge is to accurately merge these reads, which involves identifying overlaps and constructing a sequence that best represents the original RNA. The complexity lies in distinguishing between true overlaps that represent contiguous RNA sequences and erroneous overlaps due to sequencing errors or repetitive elements. The assembler must also contend with variations in gene expression levels and the presence of alternative splicing, which can complicate the reconstruction of complete and accurate transcripts.

De novo assembly predominantly relies on graph-based algorithms, with the De Bruijn graph method being particularly popular. In this approach, the reads are broken down into smaller sequences of a fixed length, known as k-mers. These k-mers are then used to construct a graph, where paths through the graph represent possible arrangements of the reads into longer sequences. The choice of k-mer length is critical; longer k-mers can help reduce errors but may miss shorter, less frequent transcripts, while shorter k-mers can capture more diversity but increase the risk of erroneous assemblies. Assemblers such as Trinity, SOAPdenovo-Trans[51], and rnaSPAdes[58] are examples of tools that employ these graph-based methods, each with unique algorithms and optimizations to handle the complexities of transcriptome assembly.

The selection of an appropriate assembly tool is crucial and depends on various factors, including the nature of the RNA-Seq data, the organism's complexity, and the specific goals of the research. Factors like computational efficiency, error tolerance, and the ability to handle alternative splicing are key considerations. Additionally, many assemblers offer the option to work with different k-mer sizes, allowing for a degree of customization based on

the specific requirements of the dataset. Researchers often compare results from multiple assemblers or k-mer settings to achieve the most comprehensive and accurate assembly possible.

After the assembly is complete, it is subjected to rigorous evaluation to assess its quality and completeness. Metrics such as N50[59] (which provides an indication of contig length) and the number of assembled transcripts is commonly used to evaluate the assembly. Tools like BUSCO assess the assembly for the presence of conserved genes, giving an insight into its completeness. The final assembly may also undergo further refinement, such as the removal of redundant sequences or the merging of fragments, to produce a coherent and representative transcriptome. This post-assembly processing is crucial in ensuring that the final product is not only comprehensive but also biologically relevant and accurate.

Post-assembly quality control

Following the intricate process of de novo transcriptome assembly, a crucial stage is the post-assembly quality control. This phase is essential for validating the accuracy and integrity of the assembled transcriptome. Despite the advancements in assembly algorithms, the output can include errors such as misassemblies, chimeric sequences, and fragmented transcripts. Therefore, quality control is vital to ensure that the assembled transcriptome is a reliable representation of the actual biological transcripts, free from artifacts that could skew downstream analyses and interpretations.

One of the primary goals of post-assembly quality control is to assess the completeness and accuracy of the assembly. Tools like BUSCO are instrumental in this process, as they compare the assembled sequences against a database of conserved genes expected to be present in all eukaryotes. High scores in BUSCO analysis indicate a comprehensive and complete assembly. Additionally, metrics such as the N50 value, which reflects the contig length where 50% of the total length of the sequences is contained, are used to gauge the continuity of the assembly.

Misassemblies and chimeric sequences[60] are common challenges in transcriptome assembly. They can arise from repetitive sequences, sequencing errors, or incorrect merging of reads. Tools like TransRate[61] and DETONATE[62] evaluate the assembly by aligning the reads back to the contigs, identifying regions with poor support or inconsistencies. This step is crucial

for detecting and rectifying erroneous assemblies, ensuring the biological credibility of the assembled sequences.

The post-assembly phase often involves refinement processes to optimize the transcriptome. This may include the removal of redundant sequences, merging of fragmented transcripts, and elimination of contigs with very low expression levels or poor alignment support. Some tools integrate multiple quality metrics and offer automated refinement suggestions, streamlining the process. These refinement steps are critical for reducing the complexity of the assembly and enhancing its utility for subsequent biological analyses.

Post-assembly quality control is not merely a technical necessity; it is a fundamental step that underpins the reliability of transcriptome-based research. A thorough quality assessment ensures that subsequent studies, such as differential expression analysis, functional annotation, and comparative genomics, are based on accurate and meaningful data. As such, this phase is integral to the overall success of RNA-Seq experiments, ensuring that the insights derived are based on a solid foundation of high-quality transcriptomic data.

Alignment and abundance estimation

The alignment process can be challenging due to the complexity of transcriptomes, which include multiple transcript variants, alternative splicing events, and a wide range of expression levels. Tools like Bowtie2[63] and STAR[64] are commonly used for this purpose, offering efficient and accurate alignment capabilities. These aligners must handle the intricacies of transcriptome data, such as reads that map to multiple locations (multi-mapping reads) or reads that span splice junctions. The choice of aligner and its configuration play a significant role in the reliability of the alignment results.

Once alignment is completed, the focus shifts to estimating the abundance of each transcript in the sample. This step is vital for understanding the expression levels of genes under different conditions. Abundance estimation is a complex task due to factors like multi-mapping reads and the need to differentiate between transcript isoforms. Tools like Salmon[65] provide solutions for these challenges. These tools estimate the number of reads originating from each transcript, accounting for the complexities of RNA-Seq data.

Pseudoalignment, a newer approach to transcript abundance estimation, offers significant advantages in terms of speed and memory efficiency. Unlike traditional alignment, pseudoalignment does not require the alignment of each read to the transcripts at the

nucleotide level. Instead, it identifies the set of transcripts that each read could have originated from based on k-mer similarity. This approach, employed by tools like Kallisto and Salmon, significantly reduces computational requirements while maintaining accuracy, making it a popular choice for large-scale RNA-Seq datasets.

Abundance estimation is more than a mere quantification step; it provides the foundational data for many downstream analyses in transcriptomics. It enables the comparison of gene expression levels across different samples or conditions, forming the basis for differential expression analysis. The accuracy of abundance estimation is critical as it directly affects the interpretation of gene expression patterns and the biological conclusions drawn from the study. Therefore, this step is integral to the success of RNA-Seq experiments, bridging the gap between raw sequencing data and meaningful biological insights.

Assembly thinning and redundancy reduction

In the aftermath of de novo transcriptome assembly, researchers are often faced with a large and complex set of sequences. This complexity necessitates the process of assembly thinning and redundancy reduction, aiming to refine the assembled transcripts into a more manageable and biologically relevant set. This step is crucial because assemblies can often include numerous redundant sequences, artifacts, or overly fragmented transcripts. Properly thinning and reducing redundancy in the assembly not only streamlines subsequent analyses but also enhances the overall accuracy and interpretability of the data.

Redundancy in transcriptome assemblies can arise from various sources, such as the assembly of highly similar paralogs, alternative splicing variants, or sequencing artifacts. Identifying and removing these redundant sequences is essential to prevent the inflation of gene counts and to ensure a more accurate representation of the transcriptome. However, this process is challenging, as it requires distinguishing between true biological variants and redundant sequences. Effective redundancy reduction not only simplifies downstream analyses like differential expression studies but also aids in better understanding the transcriptomic landscape of the organism.

Assembly thinning often involves clustering related transcripts and selecting representative sequences for each cluster. Tools like CD-HIT[66] and MMseqs2[67] are commonly used for this purpose. These tools cluster sequences based on sequence similarity and coverage, allowing researchers to either choose the longest sequence or the most representative sequence based

on consensus within each cluster. Some tools also integrate expression level data, favoring sequences with higher expression as representative of a particular transcript group.

A significant aspect of assembly thinning is dealing with transcript isoforms and paralogs. While it's important to reduce redundancy, care must be taken not to overly condense transcript variants that represent legitimate biological diversity, such as alternative splicing events. Tools like Corset and Compacta help group isoforms of the same gene, allowing for a more nuanced approach to thinning that preserves essential transcript diversity. Decisions on which isoforms to keep can be guided by factors such as expression levels, sequence completeness, and functional annotations.

Assembly thinning and redundancy reduction are not merely technical steps in the analysis pipeline; they have profound implications for the biological interpretation of the transcriptome. By reducing complexity and focusing on biologically relevant transcripts, these processes make downstream analyses like functional annotation and comparative genomics more manageable and meaningful. A carefully thinned and non-redundant assembly is a critical resource for accurate gene expression profiling and for uncovering the functional intricacies of the transcriptome.

Differential expression analysis

Differential expression analysis is a fundamental component of RNA-Seq studies, aimed at identifying genes or transcripts that show statistically significant changes in expression under different conditions. This analysis is pivotal for understanding the molecular mechanisms underlying various biological processes, such as development, response to stimuli, or disease progression. By comparing transcript abundance across different samples, researchers can pinpoint which genes are upregulated or downregulated in response to specific conditions, providing insights into the functional consequences of these expression changes.

Before undertaking differential expression analysis, it's essential to have a well-prepared dataset. This involves ensuring that the RNA-Seq data is accurately assembled and that transcript abundances are precisely estimated. The process begins with the alignment of RNA-Seq reads to a reference genome or a de novo assembled transcriptome, followed by the quantification of transcript abundances. These quantified expression levels form the basis for subsequent comparative analysis. Appropriate experimental design and replication are also crucial to account for biological variability and to ensure the reliability of the findings.

Differential expression analysis involves sophisticated statistical methods to identify significant changes in gene expression. Tools[68] like DESeq2, edgeR, and limma are commonly employed, each utilizing different statistical models to handle count data from RNA-Seq experiments. These tools account for factors such as biological variability, sequencing depth, and normalization of read counts. The choice of tool and statistical method depends on the specific nature of the data and the experimental design, with each method offering distinct advantages under different scenarios.

One of the primary challenges in differential expression analysis is distinguishing true biological changes in expression from technical noise or variability. Factors such as batch effects, sample preparation differences, and sequencing depth can influence the results. Moreover, the multiple testing problem, where the chance of finding false positives increases with the number of tests performed, is a significant concern. Careful statistical correction methods and stringent significance thresholds are employed to mitigate these issues, ensuring the robustness of the results.

The results of differential expression analysis have wide-ranging implications in biological and medical research. They can reveal key genes involved in disease mechanisms, identify potential biomarkers for diagnostics, and uncover targets for therapeutic intervention. In fundamental research, differential expression analysis contributes to our understanding of gene regulation, cellular responses, and evolutionary processes. The insights gained from this analysis form a foundation for further functional studies, such as gene knockdown or overexpression experiments, to validate and explore the roles of differentially expressed genes.

Transcriptome functional annotation

Transcriptome functional annotation is a critical step in RNA-Seq analysis that follows the assembly and quantification of transcripts. This process involves assigning biological information to the sequences, transforming raw data into a meaningful set of genes and transcripts with known or predicted functions. Functional annotation is essential for understanding the biological implications of the transcriptome data, as it links sequences to specific genes, protein functions, biological pathways, and regulatory networks. Without this crucial step, the vast amount of sequence data generated by RNA-Seq remains a largely untapped resource.

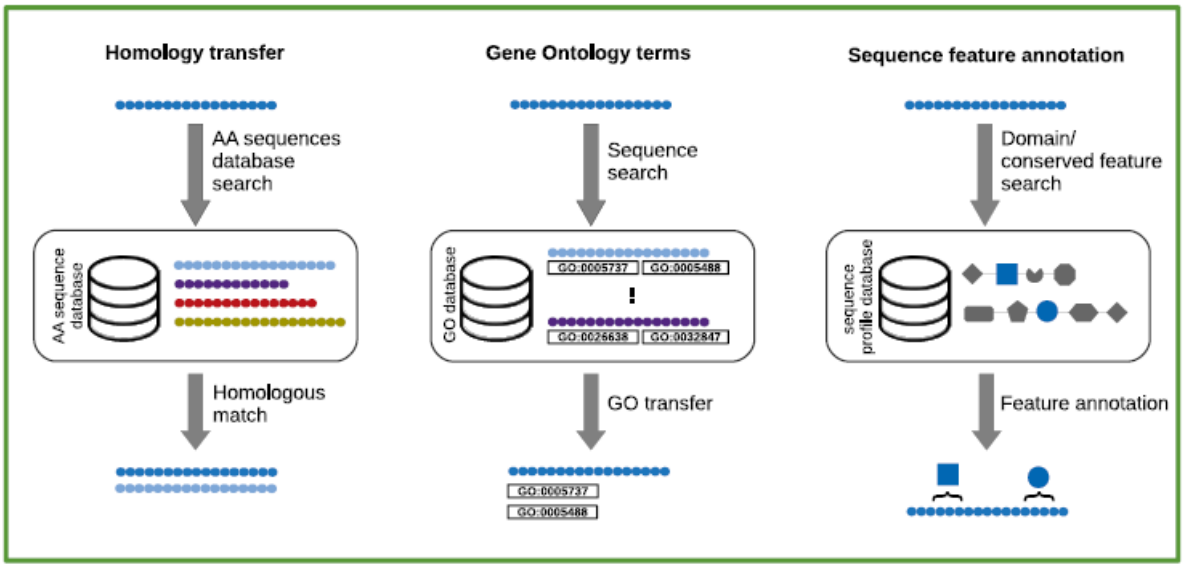


Figure 4 - Functional annotation of the transcriptome involves applying methods that attach understandable labels and functional traits to the transcripts. This process encompasses identifying homologous sequences through similarity in sequence and pinpointing assembled sequences, a process also known as homology transfer, along with domain identification.

The process of transcriptome annotation typically begins with the identification of coding regions within the RNA sequences. Tools like ORFfinder[69] and AUGUSTUS[70] are used to predict open reading frames (ORFs) and potential protein-coding sequences. However, challenges arise due to the presence of non-coding RNAs and alternatively spliced variants, which require different annotation strategies. Additionally, annotating a transcriptome from a non-model organism or one with a poorly characterized genome adds complexity, as there may be limited reference data for comparison.

Functional annotation heavily relies on homology searches against databases like NCBI's non-redundant database, UniProt[71], and Pfam[72]. These searches involve aligning transcript sequences to known sequences in the database to identify similarities that suggest common functions or evolutionary relationships. Tools such as BLAST[73] and HMMER[74] are instrumental in this process. The results from these searches can provide insights into gene function, protein domains, and potential roles in biological processes or pathways.

Beyond basic gene and protein function annotation, transcriptome annotation often involves additional layers of information. This includes mapping transcripts to known pathways using databases like KEGG[75,76] or Reactome[77], annotating regulatory elements like promoters and enhancers, and predicting RNA secondary structures. Such comprehensive annotation

provides a more complete understanding of the transcriptome's functional landscape, offering insights into the complex regulatory mechanisms governing gene expression and cellular function.

Transcriptome functional annotation is more than a data processing task; it is a gateway to numerous biological insights. Well-annotated transcriptomes facilitate a wide range of research endeavours, from comparative genomics and evolutionary studies to the discovery of novel genes and therapeutic targets. In medical research, annotated transcriptomes can reveal disease mechanisms and aid in the identification of biomarkers. In plant and agricultural sciences, they can uncover genes responsible for traits like drought resistance or yield. Thus, the thorough annotation of transcriptomes is pivotal in advancing our understanding of biology and in addressing various health and environmental challenges.

Identity assignment via homology transfer

Identity assignment via homology transfer is a fundamental component of transcriptome functional annotation. This process involves leveraging existing knowledge about genes and proteins from well-studied organisms to infer the functions of newly sequenced transcripts. The principle behind homology transfer is that sequences that are evolutionarily conserved across different species often retain similar functions. By identifying these conserved sequences, researchers can assign putative functions, biological roles, and molecular pathways to transcripts, even in the absence of direct experimental evidence.

Homology transfer typically involves sequence alignment techniques that compare the newly assembled transcripts to known sequences in established databases. Tools such as BLAST (Basic Local Alignment Search Tool) are widely used for this purpose. These tools search large databases like NCBI's non-redundant protein database, UniProtKB[78], or the Pfam protein family database to find sequences that are like the transcripts in question. The degree of similarity, or sequence homology, provides a basis for inferring the potential functions of the transcripts.

While homology transfer is a powerful tool for functional annotation, it comes with challenges and limitations. One key challenge is the distinction between orthologs (genes in different species that evolved from a common ancestral gene) and paralogs (genes related by duplication within a genome). Misidentifying these can lead to incorrect function assignment. Additionally, the method relies on the quality and comprehensiveness of existing databases;

transcripts with no close matches in these databases may remain unannotated. Furthermore, homology transfer assumes that function is conserved along with sequence, which may not always be the case, especially for rapidly evolving genes.

To maximize the effectiveness of homology transfer, it is often integrated with other annotation approaches. These may include domain prediction tools like InterProScan[79], which can identify specific protein domains and motifs, and gene ontology tools that assign transcripts to standardized categories based on biological processes, cellular components, and molecular functions. Combining homology transfer with these methods provides a more comprehensive and nuanced view of the transcriptome, capturing both broad functional categories and specific molecular activities.

Homology transfer plays a critical role in advancing our understanding of newly sequenced organisms, particularly non-model organisms with limited prior genomic information. By providing initial functional insights, it lays the groundwork for more targeted experimental investigations. In biomedical research, homology transfer can help identify potential disease-related genes and therapeutic targets. In environmental and agricultural research, it aids in the discovery of genes associated with important traits such as stress resistance or nutrient metabolism. Thus, homology transfer is a key step in translating raw sequence data into meaningful biological knowledge.

Gene ontology and pathway annotation

Gene ontology (GO)[80] and pathway annotation are critical aspects of transcriptome analysis, providing a framework to understand the functional biology of genes and transcripts identified in RNA-Seq studies. This process involves categorizing genes and transcripts into organized sets of terms and pathways that describe their roles in biological processes, cellular components, and molecular functions. By annotating the transcriptome with these standardized terms, researchers can gain insights into the underlying biological mechanisms and how different genes interact within these processes.

Gene ontology annotation starts by associating each gene or transcript with GO terms that describe its known or predicted roles. Tools like Blast2GO[81], InterProScan, and DAVID[82] are commonly used for this purpose. They perform sequence alignments and domain predictions to link transcripts with GO terms. This annotation helps in grouping genes that contribute to

similar biological functions or processes, providing a clearer picture of their collective roles in the organism's physiology and development.

In addition to gene ontology, pathway annotation is another crucial step. It involves mapping genes and transcripts to specific biochemical pathways or signal transduction pathways, elucidating their roles in complex biological networks. Databases such as KEGG, Reactome, are invaluable resources for pathway annotation. These databases provide detailed maps of metabolic and regulatory pathways, allowing researchers to place individual genes or transcripts within the broader context of cellular and organismal functions.

One of the main challenges in GO and pathway annotation is the dynamic nature of biological knowledge. As new research continually expands our understanding of gene functions and interactions, annotations must be regularly updated to reflect the latest findings. Additionally, the specificity of annotation can vary; some genes may be annotated with broad, general terms, while others may have highly specific annotations. Ensuring the accuracy and relevance of these annotations is crucial for their utility in biological research.

GO and pathway annotations are invaluable tools for interpreting the results of transcriptome studies. They enable researchers to identify potential functional changes in response to different conditions or treatments, understand disease mechanisms, and discover new targets for therapy. In broader research contexts, such as in ecology or evolution, these annotations can shed light on the adaptive functions of genes. By providing a structured and comprehensive view of gene functions and interactions, GO and pathway annotations are essential for converting raw sequence data into actionable biological insights.

Comparing transcriptome assemblies

Comparing transcriptome assemblies is a critical aspect of RNA-Seq analysis, particularly when dealing with multiple datasets, conditions, or species. This comparative analysis provides insights into the consistency and quality of different assemblies, allowing researchers to identify shared and unique transcripts across samples. Such comparisons are essential in studies involving developmental stages, tissue types, or responses to environmental stimuli. They help elucidate the differences and similarities in gene expression and regulation, offering a broader understanding of biological processes and evolutionary relationships.

The process of comparing transcriptome assemblies involves aligning one assembly against another to assess their congruence. Tools like BLAST or specialized software like ABySS[83] are used for this purpose. These tools can identify common and unique sequences, provide alignment statistics, and highlight structural variations between assemblies. The comparison can be qualitative, focusing on gene content and sequence similarity, or quantitative, evaluating aspects like transcript abundance and coverage.

One of the main challenges in comparing transcriptome assemblies is the inherent variability in RNA-Seq data. Factors such as sequencing depth, read length, and assembly algorithms can influence the outcome of the assembly, making direct comparisons complex. Additionally, variations in gene expression and alternative splicing across different conditions or species add layers of complexity to the analysis. Ensuring that assemblies are comparable in terms of quality and completeness is essential for drawing meaningful conclusions from the comparison.

Comparing transcriptome assemblies serves as a vital quality control measure. By evaluating assemblies against established references or between replicates, researchers can identify potential errors or biases in the assembly process. This comparison can also guide improvements in assembly methods, helping to refine algorithms and parameters for more accurate and comprehensive assemblies. Furthermore, it assists in validating novel transcripts or isoforms identified in an assembly, ensuring their biological relevance.

Beyond quality control, comparing transcriptome assemblies has broader implications in genomics and evolutionary biology. It allows researchers to explore gene conservation and divergence across different species or populations, contributing to our understanding of evolutionary processes. In functional genomics, such analyses can reveal the impact of genetic variations on gene expression and phenotype. Comparative assembly analysis thus plays a crucial role in advancing our knowledge of genome evolution, species adaptation, and the molecular basis of traits and diseases.

Chapter 3 - Material and Methods

Introduction

This chapter is an in-depth exploration of the techniques and strategies employed in de novo transcriptome assembly, a fundamental aspect of modern genomics when a reference genome is not available. This chapter aims to provide a comprehensive overview of the methodologies applied in this research, starting from the raw data acquisition to the complex bioinformatics analyses. The focus on de novo transcriptome assembly underscores its importance in uncovering novel insights into the genetic makeup of organisms, particularly those less studied or with significant genetic diversity.

The chapter meticulously details the methods used for sequencing and the initial processing of RNA-Seq data. It delves into the intricacies of handling high-throughput sequencing data, emphasizing the critical steps of quality assessment and the filtering of raw reads. This is essential for ensuring the reliability of the data, as the accuracy of de novo assembly is heavily contingent on the quality of the input sequences. The procedures outlined here are crucial for setting a strong foundation for the subsequent assembly process.

Subsequently, the chapter addresses the core aspect of de novo transcriptome assembly. It provides a thorough explanation of the tools and algorithms used, such as BUSCO[54], rnaSPAdes[58], TransRate[61], and DETONATE[62]. Each of these tools plays a distinct role in the assembly process, from constructing the initial transcriptome assembly to assessing its quality and completeness. The selection of these tools, their configuration, and the rationale behind their use are discussed in detail, offering insights into the decision-making process inherent in bioinformatics research.

Lastly, the chapter explores the annotation and functional analysis of the assembled transcriptome. This segment is pivotal as it transitions from mere sequence assembly to understanding the biological implications of the data. The methods for annotating the assembled sequences, including the identification of coding regions and the prediction of gene functions, are explained. This process is vital for translating the raw sequence data into meaningful biological information, which can then be used for further genetic and ecological studies.

Sample collection and RNA preparation

The analytical work began with data obtained from samples collected by the research group from the Department of Ecology and Biology at the University of Tuscia in Viterbo. Therefore, all details related to "Sample Collection and RNA Preparation" and "Library Preparation and Sequencing" are deferred to the respective sections of the scientific publications published for each studied species: *Bombina pachypus*[40], *Hyla sarda*[39], *Salamandra salamandra*[84], and *Contracaecum osculatum*[85] sp. D.

Bioinformatic pipeline and tools

Figure 5 represents the bioinformatic pipeline designed, developed, built and employed for the assembling of the *de novo* transcriptome for the non-model organisms described previously in Chapter 1.

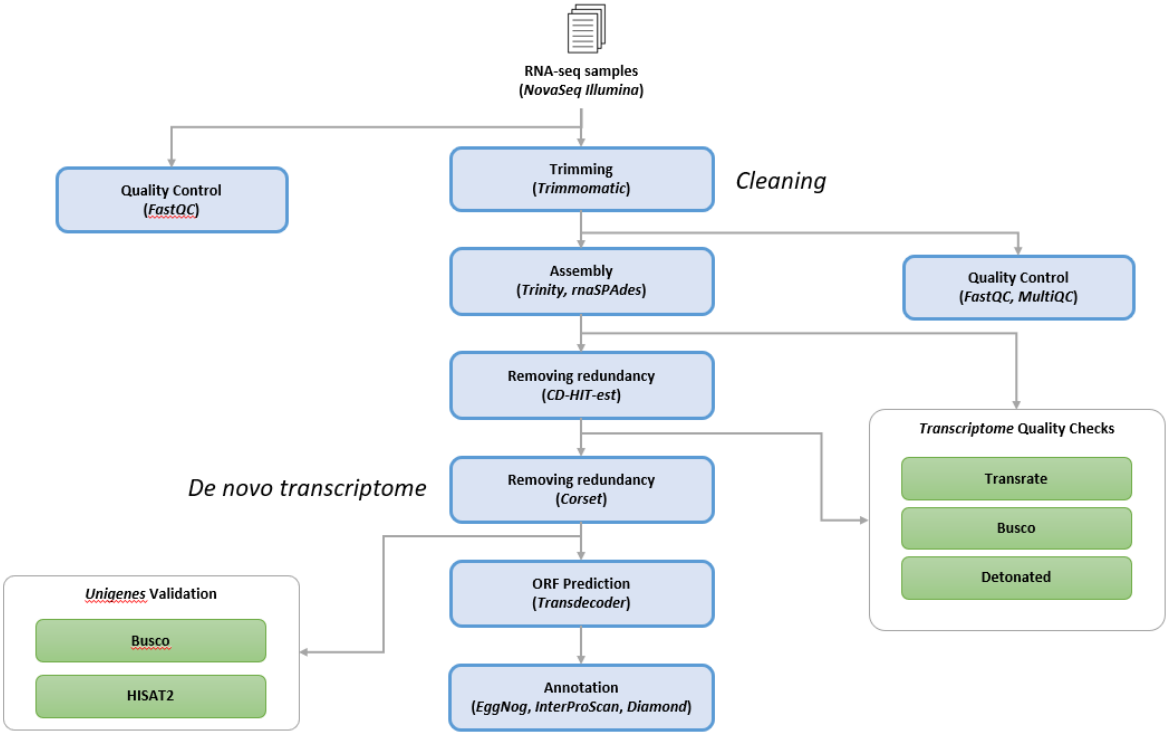


Figure 5 - This diagram represents a pipeline that has been designed, developed, built and employed for assembling, evaluating the quality of, and annotating the transcriptomes assembled de novo from non-model organisms. Modifications to this pipeline have been implemented to cater to the unique requirements of different species' sample collections.

RNA-Seq quality Controls tools

These tools play a vital role in the quality control stages of genomic and transcriptomic data analysis, guaranteeing the dependability and precision of subsequent analyses. As illustrated in Figure 5, FastQC[56] was applied to the RNA-Seq samples both prior to and following the

trimming process. Initially, it assessed the quality of the raw reads, and subsequently, it evaluated the quality of the trimmed reads to verify their condition post-cleanup.

FastQC

FastQC is designed for quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a quick and simple way to assess metrics such as the quality score distribution, sequence quality histograms, GC content, N content, sequence length distribution, and more. FastQC generates reports with graphs and tables which can help in identifying problems in the sequencing data, ensuring the data is of high enough quality for further analysis.

Version used: 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

MultiQC

While FastQC is used for analyzing individual samples, MultiQC[86] aggregates results from bioinformatics analyses across many samples into a single report. It's capable of parsing results from a variety of other bioinformatics tools, not just FastQC. MultiQC provides a global overview of the datasets, making it easier to compare the quality and output of different samples or different runs. This is particularly useful in large studies where comparing individual reports from each sample would be time-consuming and inefficient. MultiQC is capable of create images like Figure 6

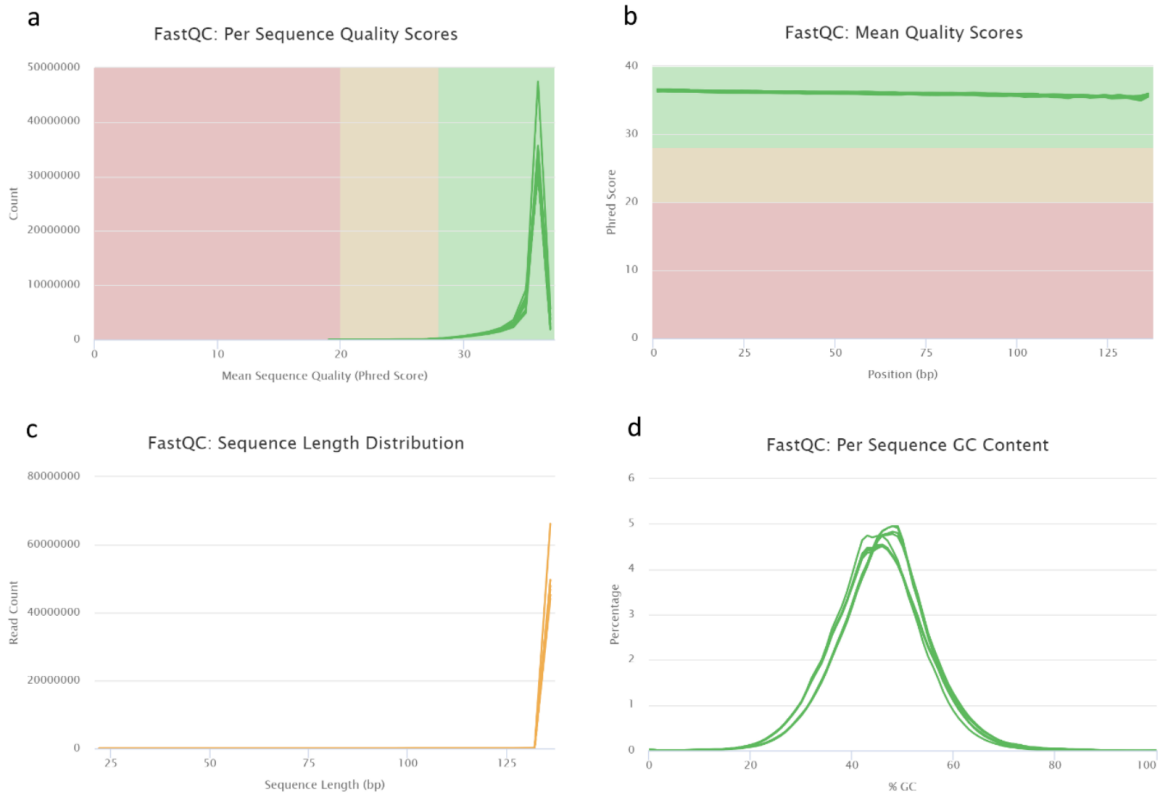


Figure 6 - Examples of images for assessing the quality of RNA-Seq data produced using FastQC in combination with MultiQC.

Version used: 1.9 (https://multiqc.info/)

Trimming tools

In RNA sequencing (RNA-Seq) and assembly, *trimming* algorithms play a crucial role in preprocessing the raw data. Here's what they typically do:

1. **Remove Low-Quality Bases:** Sequencing reads often have a decrease in quality towards the ends. Trimming algorithms identify and remove these low-quality bases to improve the overall quality of the data. This step is crucial because low-quality bases can lead to errors in downstream analyses like sequence assembly and gene expression quantification.
2. **Clip Adapter Sequences:** During the sequencing process, adapter sequences are added to the fragments of RNA to facilitate the sequencing reaction. These adapters need to be removed from the reads before analysis, as they are not part of the actual RNA sequence and can interfere with sequence alignment and assembly.

3. **Filter Out Short Reads:** After trimming low-quality bases and adapters, some reads may become very short. These short reads may not provide reliable information and can complicate assembly and alignment processes. Trimming algorithms often filter out these reads.
4. **Remove Contaminants**: In some cases, trimming tools can also identify and remove contaminant sequences, such as ribosomal RNA (rRNA) sequences, which are often not the focus of RNA-Seq studies.
5. **Poly-A/Tail Trimming:** For certain types of RNA-Seq data, such as those involving mRNA, trimming algorithms can be used to remove poly-A tails (a string of adenine nucleotides) or their complements, as these are also not informative for most analyses.

By performing these tasks, as shown in Figure 7, trimming algorithms help in ensuring that the RNA-Seq data is of high quality and suitable for accurate and efficient downstream analysis, such as assembly, alignment, and quantification.

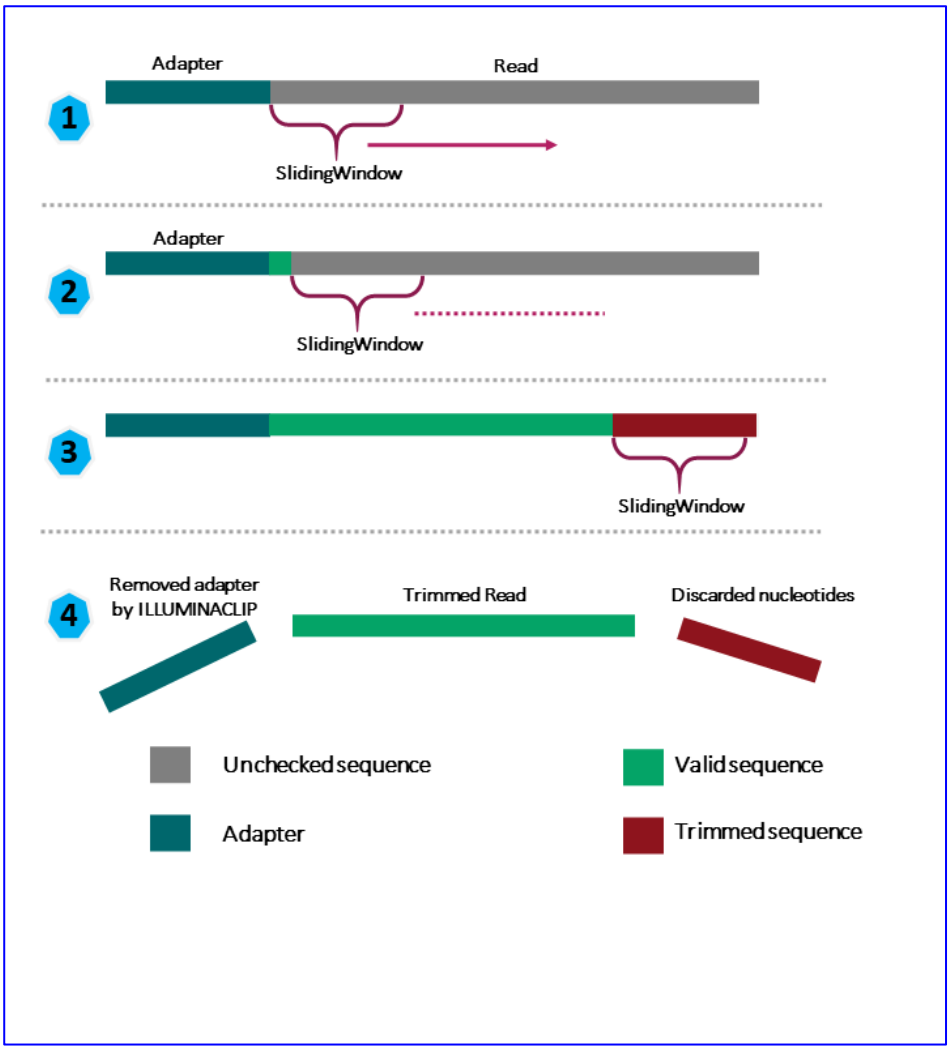


Figure 7 - Diagram illustrating the trimming procedure used for processing RNA-Seq data. The process involves the removal of adapter sequences typically present at the beginning of the raw reads, as well as the trimming of low-quality or unwanted nucleotides from the end of the reads. This ensures that only high-quality, adapter-free sequences are retained for subsequent analysis steps, enhancing the accuracy of the RNA-Seq data interpretation and downstream genomic analyses.

Trimmomatic

For the trimming phase, Trimmomatic[87] has been used. Trimmomatic was useful for:

1. **Data Quality Improvement**: The quality of sequencing reads often decreases towards the end. Therefore, trimming the low-quality bases can significantly enhance the outcomes of the analyses.
2. **Adapter Removal**: Adapters are used during the library preparation for sequencing and may occasionally be sequenced along with the reads. These adapter sequences are irrelevant for most analyses and can introduce artifacts, making their removal crucial.
3. **Handling Paired-End Reads**: Trimmomatic is capable of processing paired-end reads, ensuring that the pairs remain synchronized after the trimming process.

Example of how this tool works is represented in Figure 7. For the pipeline in Figure 5, we always used the options SLIDINGWINDOW: 4: 15, MINLEN: 36 and HEADCROP: 13.

Version used: 0.39 (http://www.usadellab.org/cms/?page=trimmomatic)

Assembly tools

Assembly tools play a crucial role in the bioinformatic pipelines, especially for tasks like RNA-Seq analysis. Their primary function is to piece together sequences from fragmented DNA or RNA reads generated by high-throughput sequencing technologies. A general overview of what they do:

1. **Sequence Assembly:** The core task of assembly tools is to construct a contiguous sequence (known as a contig) from overlapping, fragmented sequences known as reads. This process is akin to solving a jigsaw puzzle where each read is a piece of the puzzle.
2. **Dealing with High-Throughput Data:** Modern sequencing technologies produce a massive volume of short sequences. Assembly tools efficiently manage and process these large datasets to create longer sequences that represent the original DNA or RNA source.

3. **Handling Errors and Variability:** Sequencing technologies are prone to errors. Assembly tools incorporate algorithms to correct these errors, differentiate between true variants and errors, and deal with ambiguities.
4. **De Novo Assembly:** When there is no reference genome available (common in non-model organisms as the case of the species studied during this work), de novo assembly tools build the sequence from scratch, relying solely on the information contained in the reads.
5. **Reference Assembly:** If a closely related reference genome is available, assembly tools can align the new reads to this reference, making the assembly process more accurate and less computationally intensive.
6. **Contig Linking:** Longer sequences, known as scaffolds, can be created by joining contigs. This often involves using additional information like mate-pair sequencing, which provides data on how far apart two reads are in the original DNA.
7. **Quality Assessment:** Post-assembly, these tools often provide metrics to assess the quality of the assembled sequence, such as the length of contigs, coverage depth, and the presence of gaps. In our case, we used more sophisticated and dedicated algorithms to evaluate the quality assessment for the de novo assemblies generated.
8. **Optimization for Organism-Specific Challenges:** Certain organisms or sample types may present unique challenges (e.g., high GC content, repetitive regions). Some assembly tools are specifically optimized to handle these peculiarities more effectively.

Each assembly tool may have its unique strengths and be better suited for specific types of data or research goals. Therefore, choosing the right tool is crucial for accurate and efficient assembly in genomics research, for this reason we didn't always use the same algorithm, but we preferred to generate different assemblies with different algorithms and use which with the best in quality, comparing different parameters, e.g. the N50 value, number and length of the sequences. Following assembling algorithms were used for this work:

Trinity

Trinity[50] is a highly regarded software tool in the field of bioinformatics, specifically designed for the de novo reconstruction of transcriptomes from RNA-Seq data. It is particularly useful in scenarios where a reference genome is not available, such as in studies

involving non-model organisms. Trinity is recognized for its robustness and efficiency in assembling RNA-Seq reads into longer sequences, known as transcripts, which are crucial for further analysis such as gene expression studies, functional annotation, and discovery of novel genes. Two key aspects of Trinity are the *Comprehensive Assembly Process* and the *Adaptability and Sensitivity*:

1. **Comprehensive Assembly Process:** Trinity assembles transcriptome data through a process that involves three main stages: Inchworm, Chrysalis, and Butterfly. Inchworm starts by constructing longer sequences, or contigs, from the overlapping regions of short RNA-Seq reads. Chrysalis then groups these contigs based on shared components and constructs de Bruijn graphs for each group. The final stage, Butterfly, processes these graphs to tease out the distinct transcripts (isoforms) and resolves potential complications such as alternative splicing or artifacts from sequencing errors. This approach allows Trinity to effectively reconstruct a comprehensive set of transcripts, making it particularly adept at handling complex transcriptomes with a high degree of gene isoform diversity.
2. **Adaptability and Sensitivity:** Trinity is designed to adapt to various types of RNA-Seq data, including those from different sequencing platforms and experimental setups. It is sensitive enough to detect rare transcripts and capable of handling large datasets, which is particularly important given the ever-increasing volume of data generated by modern high-throughput sequencing technologies. Trinity's ability to work with data from single or paired-end reads, as well as its proficiency in assembling full-length transcripts for a majority of genes, makes it a versatile tool for transcriptome assembly.

Used version: https://github.com/trinityrnaseq/trinityrnaseq/wiki

rnaSPAdes

rnaSPAdes[58] is a specialized tool for de novo transcriptome assembly, developed as an extension of the SPAdes genome assembler, which is widely recognized in the field of genomics. It is specifically tailored to handle RNA sequencing (RNA-Seq) data, making it a valuable resource for researchers working on transcriptome projects, particularly when no reference genome is available. Here are two key aspects of rnaSPAdes that highlight its functionality and application:

1. **Specialized Algorithm for RNA-Seq Data:** Unlike general genome assemblers, rnaSPAdes is designed to address the unique challenges presented by RNA-Seq data. RNA-Seq data typically have variable coverage, with certain regions being overrepresented due to the expression level differences across genes. rnaSPAdes employs an algorithm that accounts for these coverage discrepancies, effectively assembling transcripts across a wide range of expression levels. This includes the ability to handle both highly expressed genes and those present at low levels, which might be missed by other assemblers. Additionally, rnaSPAdes is adept at dealing with the intricacies of transcriptome data, such as alternative splicing events, which can produce multiple transcript variants from a single gene.
2. **Versatile and User-Friendly:** rnaSPAdes supports various types of input data, including single-end, paired-end, and mate-pair reads, making it versatile for different sequencing protocols. It's also compatible with data from different sequencing platforms like Illumina, IonTorrent, and PacBio. This flexibility is a significant advantage for researchers working with diverse datasets. Moreover, rnaSPAdes is user-friendly, with straightforward command-line operations and clear documentation, facilitating its use by both experienced bioinformaticians and researchers new to transcriptome assembly. Its integration into the SPAdes software package, which is well-regarded for its efficiency and accuracy in genome assembly, further enhances its reliability and appeal in the scientific community.

rnaSPAdes stands out as a robust and efficient tool for transcriptome assembly, capable of handling the complex nature of RNA-Seq data while remaining accessible to a broad range of users. Its specialized algorithms provide high-quality assemblies, making it an essential tool for genomics research, particularly in studies involving non-model organisms or when exploring novel transcripts and gene isoforms.

Used version: v.3.14.1 (https://cab.spbu.ru/software/spades/)

Merging tools

These tools are specifically designed to tackle the challenge of short, fragmented sequences, a common output of rnaSPAdes and DNA sequencing processes. They function by identifying overlaps and similarities between these fragmented sequences, known as reads, and adeptly combine them to form longer, contiguous sequences (contigs). This process is not only vital for the construction of a genome or transcriptome in de novo assembly, where no reference

genome is available, but also enhances the accuracy and integrity of reference-guided assemblies.

The versatility of merging tools becomes particularly evident when dealing with data from various sequencing technologies, such as Illumina, PacBio, and Oxford Nanopore. Each of these technologies has its unique characteristics in terms of read length and error profiles, and merging tools have evolved to accommodate these differences efficiently. By processing such diverse data, these tools contribute significantly to the extension of contigs and scaffolds, thus delivering a more complete and accurate representation of the genome or transcriptome sequences under study.

Trans-ABySS:

Trans-ABySS[88] is a specialized bioinformatics tool designed for the effective assembly of RNA and DNA sequencing data, particularly adept at handling data generated from high-throughput sequencing technologies. Developed as an extension of the ABySS[83] (Assembly By Short Sequences) framework, Trans-ABySS is tailored to address the challenges posed by sequencing projects that generate a large number of short reads. It operates by first fragmenting the sequence data into smaller, manageable pieces (k-mers), and then reconstructing the original sequence by finding overlaps between these k-mers. This method allows Trans-ABySS to create longer contiguous sequences (contigs) from short, fragmented data, making it particularly useful for de novo assembly where no reference genome is available.

Trans-ABySS, in its role as a merger tool, plays a crucial part in refining and enhancing the quality of transcriptome and genome assemblies. This software stands out for its ability to effectively combine and consolidate multiple de novo assemblies, each generated using different k-mer sizes. By doing so, Trans-ABySS can capture a broader range of sequence complexities and variances that might be missed when using a single k-mer size. This merging process is vital for achieving a more comprehensive and accurate assembly, particularly in scenarios dealing with complex genomes or transcriptomes that exhibit high variability and intricacies, such as those with extensive repetitive regions or significant gene isoform diversity.

The strength of Trans-ABySS as a merger tool lies in its algorithmic design, which not only facilitates the integration of assemblies but also optimizes the process to handle the large and varied datasets typical in high-throughput sequencing projects. This optimization is crucial

for managing the computational challenges posed by large-scale genomic data. By reconciling assemblies from multiple k-mers, Trans-ABySS ensures that the final merged assembly is both of high quality and representative of the true biological complexity of the sample. This makes Trans-ABySS an indispensable tool in the toolbox of genomics and transcriptomics researchers, aiding in unraveling the intricate details of genetic information with greater accuracy and efficiency.

Used version: 2.0.1 (https://www.bcgsc.ca/resources/software/trans-abyss)

Removing Redundancy tools

Removing redundancy tools in bioinformatics are essential in the analysis of high-throughput sequencing data, particularly in RNA sequencing projects. These tools address the challenge of managing the vast amounts of data produced by these technologies, which often include a considerable number of duplicate or highly similar sequences. By identifying and grouping these redundant sequences, the tools streamline the dataset, enhancing its quality for more accurate downstream analyses. This process not only simplifies the data but also significantly improves the efficiency of computational analyses by reducing the volume of data to be processed.

The operation of these tools involves advanced computational methods to compare sequences within a dataset, identifying those that are identical or exhibit a high degree of similarity. Techniques like sequence alignment and heuristic algorithms are employed to cluster similar sequences. Each cluster is then represented by a single sequence, effectively reducing the total number of sequences in the dataset. Users can set a similarity threshold to define the extent of similarity for sequences to be considered redundant. This flexibility allows the tools to be customized according to the specific needs of the research.

In practical applications, these tools are invaluable for tasks such as transcriptome assembly, where they help in constructing transcripts from short reads and identifying unique gene transcripts. They are also crucial in genome annotation, ensuring that each genomic feature is represented uniquely. Overall, the utilization of redundancy removal tools is a key step in the bioinformatics workflow, pivotal in ensuring that the data analyzed is not just manageable but also accurate and reliable for scientific research.

CD-Hit-Est

CD-HIT-est[66] is a widely used bioinformatics tool designed for clustering DNA sequences to reduce redundancy and improve the efficiency of downstream analyses. Its fundamental

purpose is to group sequences that are like each other based on a user-defined similarity threshold.

The way CD-HIT-est works is straightforward yet powerful. It sorts sequences by length, keeping the longest sequence as a representative of the cluster. Subsequent sequences are then compared to these representatives. If a sequence shares a similarity above the threshold with a representative, it is grouped into that cluster and not considered for further representative status. This process continues until all sequences have been processed. The similarity threshold, typically a percentage, determines how closely the sequences within a cluster are related. The efficiency of CD-HIT-est arises from its use of a heuristic algorithm that speeds up the comparison process, allowing it to handle large datasets effectively.

The primary output of CD-HIT-est is a set of clusters, each with a representative sequence, significantly reducing the number of sequences in datasets without losing the diversity of the original data. This clustering is crucial in transcriptome analysis where redundancy and repeat sequences are common, potentially leading to errors in downstream analyses like assembly, annotation, and expression analysis.

Used version: 4.8.1 (https://sites.google.com/view/cd-hit)

Corset

Corset[89], on the other hand, is another essential tool used in bioinformatics, particularly in the context of RNA-Seq data. It clusters contigs (sequences assembled from RNA-Seq reads) that likely originate from the same transcript, thereby helping in reducing redundancy and improving the interpretability of the data.

Corset operates by first aligning RNA-Seq reads to the assembled contigs. It then utilizes the information about which reads map to which contigs to group contigs into clusters. The rationale behind this is that reads from the same transcript will map to contigs that are likely parts of that transcript. Therefore, by analyzing the patterns of shared reads among contigs, Corset can deduce which contigs are part of the same transcript.

The tool uses a hierarchical clustering algorithm, which starts by treating each contig as a separate cluster. It then progressively merges clusters based on the similarity of their read mapping profiles. The decision to merge clusters is based on statistical models that estimate the likelihood of observing the shared read patterns under the hypothesis that the clusters are from the same or different transcripts.

The output of Corset is a set of clusters with reduced redundancy, where each cluster represents a group of contigs likely originating from the same transcript. This clustering facilitates more accurate downstream analyses like differential expression analysis, as it reduces the noise and complexity of the data by consolidating contigs that represent the same transcript.

Used version: 1.0.6 (https://github.com/Oshlack/Corset/wiki)

Assembly evaluation tools

In the field of bioinformatics, transcriptome quality check tools are essential for evaluating the accuracy and completeness of de novo transcriptome assemblies. These tools are particularly important when assembling transcriptomes without a reference genome, a process that can be prone to various errors such as missassemblies, chimeric transcripts, and incomplete or fragmented transcripts. Quality check tools like BUSCO54, TransRate[61], and DETONATE[62] provide crucial metrics and insights to assess and ensure the integrity of these assemblies.

TransRate is a tool that assesses the quality of a transcriptome assembly by analyzing the contiguity and coverage of the assembly. It calculates metrics such as the N50 value, which reflects the length of contigs, and the proportion of reads that correctly map back to the assembly. These metrics are indicative of the assembly's quality, highlighting potential issues like misassemblies or fragmentations. BUSCO (Benchmarking Universal Single-Copy Orthologs) offers another layer of assessment by comparing the assembly against a conserved set of genes known to be present in nearly all eukaryotic organisms. This comparison helps to determine the completeness of the transcriptome assembly in terms of capturing essential gene content, providing a measure of how representative the assembly is of the organism's transcriptome.

DETONATE is another valuable tool for quality assessment, employing a reference-free evaluation method based on a probabilistic model that depends solely on the assembly and the RNA-Seq reads used to construct it. This tool provides a comprehensive analysis of potential errors in the assembly process, offering insights into the overall quality of the assembled transcriptome. Together, these tools – TransRate, BUSCO, and DETONATE – play a crucial role in the de novo transcriptome assembly process. They provide researchers with quantitative evaluations and diagnostics, guiding them in refining the assembly and ensuring that the final transcriptome accurately represents the genomic expression of the

organism under study. This level of quality assessment is pivotal for subsequent analyses such as gene expression studies, annotation, and functional genomics, making these tools indispensable in the field of transcriptomics.

All the three tools previously mentioned provide detailed insights into the quality and integrity of the assembly helping to refine and improve the assembled transcriptomes, ensuring that subsequent analyses, such as gene expression studies or functional annotations, are based on reliable and accurate data.

TransRate

TransRate[61] is a software tool used for the quality assessment of de novo transcriptome assemblies, which are the constructions of mRNA transcript sequences from RNA-Seq data without a reference genome. In the context of de novo assembly, TransRate plays a crucial role in evaluating the quality and accuracy of the assembled transcriptome.

The tool works by analyzing the assembled sequences (contigs) against the RNA-Seq reads used to generate them. It provides a set of metrics that help in identifying errors in the assembly, such as chimeras, structural errors, or misassemblies. These metrics include the TransRate Assembly Score, which is an overall measure of assembly quality, and the TransRate Optimal Score, which indicates the proportion of the assembly that is of good quality. The tool also assesses the proportion of 'good contigs' in the assembly, reflecting how many of the assembled sequences are supported by the RNA-Seq reads.

Used version: 1.0.3 (https://hibberdlab.com/transrate/).

Busco

BUSCO (Benchmarking Universal Single-Copy Orthologs)[54] is a tool used to assess the completeness and quality of genome assemblies, annotated gene sets, and transcriptomes in a quantitative manner. In the context of de novo transcriptome assembly, BUSCO is particularly valuable for evaluating the quality of assemblies when no reference genome is available.

BUSCO works by comparing the assembly to a set of highly conserved genes that are expected to be present as single-copy orthologs in nearly all organisms within a specified taxonomic group. These genes serve as a standard measure for completeness. The tool reports the proportion of these conserved genes that are complete, fragmented, or missing in the given assembly. This provides an insight into how well the assembly captures the expected

gene content, thus indicating its completeness and accuracy. A high proportion of complete BUSCOs suggests a comprehensive and accurate assembly, whereas many fragmented or missing BUSCOs may indicate gaps or errors in the assembly.

Used version: 4.1.4 (https://busco.ezlab.org/).

Detonate

DETONATE (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation)[62] is a software tool used for evaluating the quality of de novo transcriptome assemblies. This tool is particularly useful in scenarios where no reference genome is available to validate the assembled transcriptome. DETONATE focuses on assessing the assembly using the RNA-Seq data that was used for the assembly itself.

The core functionality of DETONATE is to provide a probabilistic model that evaluates the transcriptome assembly based solely on the assembly and the RNA-Seq reads used to construct it. This is achieved through a reference-free method, which means it does not rely on any external reference genome for assessment. DETONATE calculates a score for the assembly that reflects its overall quality, taking into account factors such as the correctness of the assembled sequences, the coverage of the reads, and the representation of the RNA-Seq data in the assembly.

Used version: 1.11 (https://github.com/deweylab/detonate).

ORF Prediction tools

Open Reading Frame (ORF) prediction tools are vital in bioinformatics for identifying potential protein-coding regions in a genomic or transcriptomic sequence. An ORF is a segment of DNA or RNA that potentially encodes a protein, marked by a start codon at the beginning and a stop codon at the end. These tools are designed to scan sequences for these codons and delineate the regions that could encode proteins. The most straightforward approach in ORF prediction involves scanning the sequence for start and stop codons across different reading frames – each sequence has three possible reading frames per strand, owing to the triplet nature of codons.

The more advanced ORF prediction tools utilize complex algorithms and additional genomic information to enhance prediction accuracy. They may incorporate statistical models that consider codon frequency biases, known gene structures, and regulatory elements like promoters and ribosome binding sites that suggest true protein-coding regions. Some tools

even use machine learning techniques, trained on well-annotated genomes, to improve their predictive capabilities. This integration of various computational methods allows for a more robust and accurate identification of ORFs.

The significance of ORF prediction extends beyond mere identification of coding regions. The output from these tools, which includes the location, length, and potential amino acid sequence of each ORF, is instrumental in gene annotation, functional genomics, and proteomics. By predicting ORFs, researchers can infer the function of unknown genes, study gene expression patterns, and understand the genetic basis of various biological processes. This makes ORF prediction a foundational step in genomic research, crucial for unraveling the complexities of genetic information.

TransDecoder

TransDecoder[90-92] is a specialized bioinformatics tool designed for identifying candidate coding regions within transcript sequences, such as those obtained from RNA sequencing (RNA-Seq) experiments. It is particularly useful in the analysis of non-model organisms or in situations where a reference genome is not available. TransDecoder focuses on extracting the most likely protein-coding sequences from a given set of transcript sequences, which is a critical step in understanding the functional elements of a genome and the proteins it encodes.

The tool operates by first scanning the input transcript sequences for open reading frames (ORFs) that have the potential to encode proteins. TransDecoder looks for ORFs that contain a start codon, followed by a sequence that is long enough to be a plausible protein-coding region, and then a stop codon. This process is conducted across all six potential reading frames for each transcript (three in the forward direction and three in the reverse direction). To enhance the accuracy of its predictions, TransDecoder also employs heuristics and statistical models. It considers factors like the length of the ORF, the presence of a complete and intact coding region, and the similarity of the predicted protein sequence to known proteins.

The output of TransDecoder includes a set of candidate protein-coding sequences along with their corresponding amino acid translations. These results are invaluable for subsequent analyses, such as functional annotation, comparative genomics, and studies of gene expression and regulation. By providing a means to identify likely protein-coding regions in transcriptomic data, TransDecoder enables researchers to delve deeper into the understanding

of gene functions and their roles in various biological processes, especially in the context of organisms where genomic resources are limited.

Used version: 5.7.1 (https://github.com/TransDecoder/TransDecoder).

Mapping Tools

Mapping tools in bioinformatics, crucial for aligning sequencing reads to reference sequences, include genome aligners, transcriptome aligners, and specialized tools like HISAT2[93]. These tools are essential in genomic and transcriptomic analyses, enabling researchers to locate the origin of each read within a reference genome or transcriptome. This alignment is fundamental for various downstream applications, including variant detection, gene expression studies, and genomic rearrangements analysis. An example of RNA-Seq mapping is illustrated in Figure 8.

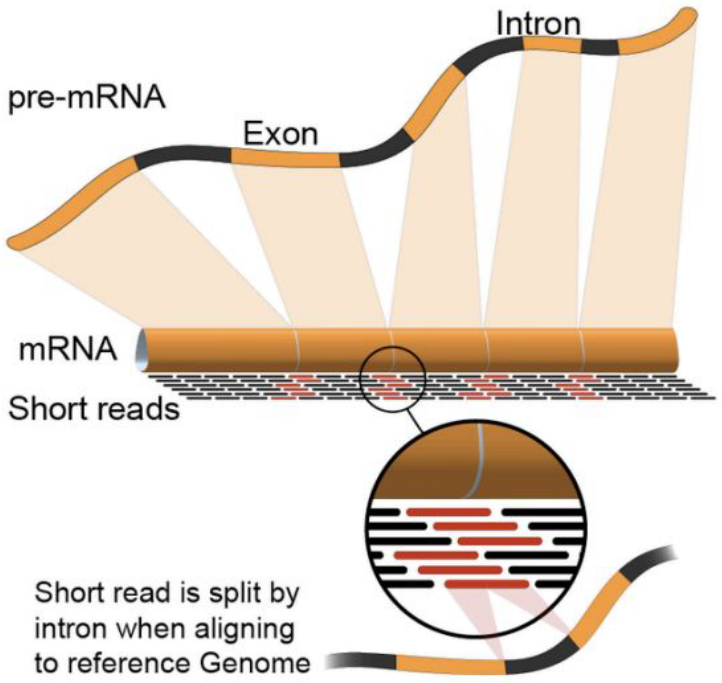


Figure 8 - Figure depicting the mapping of RNA-Seq data onto the genome or transcriptome. This process aligns RNA sequences from high-throughput RNA-Seq to a reference sequence, revealing how these sequences match specific genomic regions. This essential step is key in identifying active transcription areas, quantifying gene expression, and understanding genomic functions in biological processes. It aids in uncovering gene expression patterns and discovering new transcripts.

The process involves aligning short DNA or RNA sequencing reads, obtained from sequencing technologies to a longer reference sequence. This task is challenging due to potential errors in sequencing reads and variations in the reference sequence. Mapping tools use advanced algorithms to accommodate mismatches, insertions, deletions, and handle other

complexities. They offer parameters for adjusting alignment conditions, such as mismatch tolerance and gap handling. Genome aligners like BWA and Bowtie are tailored for aligning reads to a genome, commonly used in resequencing and variant analysis. Transcriptome aligners like HISAT2, TopHat and STAR focus on RNA-Seq data, aligning reads to a transcriptome and considering gene splicing.

HISAT2

HISAT2[93] is a highly efficient and advanced mapping tool specifically designed for aligning RNA sequencing (RNA-Seq) reads to a reference genome. It stands out for its novel indexing architecture, which allows it to handle large genomes like the human genome with remarkable speed and reduced memory requirements. HISAT2 is particularly adept at dealing with spliced alignments, a critical feature for RNA-Seq data, as it can accurately map reads across exon-exon junctions. This capability makes HISAT2 an invaluable tool for studying gene expression and alternative splicing events, which are essential for understanding complex gene regulation mechanisms and the diversity of protein products in eukaryotic cells. By efficiently aligning reads to both the genome and known splice sites, HISAT2 facilitates the identification of novel splice variants and provides insights into the transcriptional landscape of cells.

Furthermore, HISAT2's versatility extends to various applications in genomic research, including differential gene expression analysis, variant calling, and transcriptome assembly. Its ability to handle both single-end and paired-end reads, along with its compatibility with various sequencing platforms, contributes to its wide applicability in genomic studies. The output from HISAT2, typically in standard SAM or BAM formats, includes detailed alignment information that can be seamlessly integrated into downstream analytical pipelines. The tool also generates metrics that help assess the quality of the alignments, providing researchers with valuable data for interpreting their results. This combination of speed, efficiency, and accuracy makes HISAT2 a tool of choice for researchers working with RNA-Seq data, especially in projects involving large and complex genomes.

Used version: 2.2.1 (https://daehwankimlab.github.io/hisat2/).

Annotation tools

In the realm of bioinformatics, software tools for de novo transcriptome annotation are crucial for deciphering the functional aspects of transcript sequences obtained from organisms without a reference genome. These tools are adept at analyzing RNA sequences to

identify and label various genetic components like expressed genes, splice variants, untranslated regions (UTRs), and potential coding sequences. This annotation is vital for understanding the transcriptomic landscape of an organism, offering insights into gene expression patterns, regulatory mechanisms, and the functional dynamics of genes in various biological processes, particularly in non-model organisms or novel species.

The functionality of de novo transcriptome annotation tools involves sophisticated algorithms that sift through transcript sequences to identify features such as open reading frames (ORFs), potential coding regions, and consensus splice sites. This process often integrates sequence homology searches, comparing the transcript sequences to known genes and proteins in comprehensive databases to predict functions and relationships. This comparative approach is especially significant in de novo transcriptome studies, as it allows for functional inference in the absence of a reference genome, providing a basis for identifying gene orthologs, functional motifs, and evolutionary relationships.

De novo transcriptome annotation software thus serves as an invaluable tool for researchers delving into the transcriptomes of unexplored species, shedding light on gene functions, expression levels, and the molecular pathways involved in specific biological conditions. By enabling the annotation and functional prediction of transcripts in newly sequenced organisms, these tools pave the way for groundbreaking discoveries in genetics, evolutionary biology, and molecular medicine. They transform raw transcriptomic data into a rich source of biological information, facilitating a deeper understanding of the underlying genetic architecture and its role in the organism's life processes.

InterProScan

InterProScan[79] is a tool designed for the functional analysis of protein sequences. It plays a crucial role in characterizing and annotating proteins by identifying protein families, domains, and functional sites. InterProScan achieves this by integrating predictive models and signature recognition algorithms from various member databases into one powerful tool, offering a broad perspective on protein function.

The primary function of InterProScan is to scan protein sequences against a database that compiles protein signatures from multiple resources, including Pfam, PROSITE, SMART, TIGRFAMs, and others. Each of these resources contributes unique information about protein families, domains, and functional sites, making InterProScan an aggregation of diverse protein characteristics. The tool identifies specific patterns or motifs in a protein

sequence that match known signatures, thus inferring possible biological functions. This approach allows for the identification of evolutionary conserved elements within proteins, offering insights into their function, structure, and evolution.

The output from InterProScan is a detailed annotation of the input protein sequence, including the identification of any matches to known protein families, domains, active sites, and other relevant features. These annotations provide critical information for understanding the roles of proteins in various cellular processes, their potential interactions, and their involvement in pathways. InterProScan is thus an invaluable resource for researchers in fields like proteomics, structural biology, and drug discovery, where understanding the functional aspects of proteins is key. By offering comprehensive insights into protein characteristics, InterProScan facilitates the exploration of protein functions and enhances our understanding of biological systems at the molecular level.

Used version: 5-50-84.0 (https://interproscan-docs.readthedocs.io/en/latest/index.html).

Diamond

Diamond[95,96] is a high-performance bioinformatics software tool specifically designed for the rapid alignment of large sets of protein and translated DNA sequences against a protein reference database. It stands out for its exceptional speed, which significantly surpasses that of traditional tools like BLASTX and BLASTP, without compromising the accuracy of the results. Diamond's speed advantage makes it particularly useful in processing the vast amounts of data generated by modern high-throughput sequencing technologies.

The primary function of Diamond is to perform sequence alignment and homology search, which are crucial in the fields of genomics and proteomics. It aligns short DNA sequences (reads) or protein sequences to a large protein database, identifying similarities and inferring potential functions based on known proteins. This process is essential for tasks such as functional annotation of genes and proteins, metagenomic analysis, and studying evolutionary relationships among proteins. Diamond operates using a double-indexing algorithm, which allows for fast and efficient matching of sequences. This innovative approach significantly reduces the computational resources and time required for sequence alignment, making it feasible to handle extremely large datasets.

Diamond's output includes detailed alignment information, providing insights into the sequence similarities and potential functional relationships between the query sequences and those in the reference database. This information is pivotal for researchers in understanding

the roles of genes and proteins in various biological processes, exploring the diversity of life at the molecular level, and contributing to fields like comparative genomics, functional genomics, and metagenomics. Diamond's efficiency and accuracy make it a valuable tool in the bioinformatics toolkit, particularly in an era where the volume of genomic and proteomic data is growing exponentially.

Used version: 2.0.15 (https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/diamond/).

Development of custom tools

The thesis work was not limited to the use of pre-existing codes. For the submission of the generated de novo transcriptomes, it was necessary to develop new Python codes/tools for:

1. Removing sequences shorter than 200 nucleotides.
2. Eliminating sequences containing more than 10% of "N".
3. Removing sequences considered contaminated by NCBI (National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/) and EBI (European Bioinformatics Institute, https://www.ebi.ac.uk/).
4. Appropriately renaming the header of each sequence.
5. Processing the result files from InterProScan and Diamond to extract the necessary data for creating Venn diagrams, histograms, etc.
6. Performing parsing and merging of multiple files for the analysis of differentially expressed genes.
7. Enriching files with information about genes by calling appropriate web services and extracting the necessary data from the responses obtained.

Code has been developed using Python language version 3.11.0 (https://www.python.org/) and Visual Studio Code (https://code.visualstudio.com/), a cross-platform IDE (Integrated Development Environment).

Chapter 4 - Results and Discussion

Introduction

The cornerstone of our genetic exploration, the de novo transcriptome assembly and quality assessment, was a critical phase in our journey to unravel the genomic secrets of *Bombina pachypus*[40], *Hyla sarda*[39], *Salamandra salamandra*[86], and *Contracaecum osculatum*[87] *Sp. D*. This stage was characterized by the intricate assembly of RNA sequences into a comprehensive transcriptome, a task that was uniquely challenging due to the non-model status of our subjects.

Utilizing next-generation sequencing data, we embarked on the complex process of assembling the transcriptome for each species. Using the bioinformatics pipeline shown in Figure 5, and extensively detailed in Chapter 2, "Material and Methods", assembling of de novo transcriptomes involved the alignment and merging of short RNA sequences (reads) into longer contiguous sequences (contigs), which were then further joined to form complete transcripts. The diversity in our subject species, ranging from amphibians to parasites, introduced distinct challenges in terms of sequence variability and complexity. Our methodology was thus carefully adapted to accommodate these differences, ensuring the accurate and efficient assembly of each species' transcriptome.

Ensuring the quality and reliability of each assembled transcriptome was paramount. A rigorous quality assessment protocol has been employed, involving various metrics such as N50 length, contig size distribution, and GC content analysis, to evaluate the completeness and accuracy of the assemblies. Additionally, we conducted thorough checks for potential contamination and artifacts, a crucial step given the environmental and host-derived samples, particularly in the case of *Contracaecum osculatum*. This comprehensive evaluation ensured that our transcriptomes were of high fidelity, serving as a reliable foundation for further analysis and interpretation.

Through this meticulous process of de novo transcriptome assembly and quality assessment, we were able to construct detailed genetic maps for each species, overcoming the limitations imposed by their non-model organism status. This groundbreaking work not only paves the way for deeper genetic and ecological insights but also contributes significantly to the broader field of evolutionary biology and conservation genetics.

Is important to underline that this thesis work primarily concentrated on generating genomic resources, specifically de novo transcriptomes, utilizing the available data from each sample collection, and not on the sample collecting or RNA extraction and preparation. Samples were gathered from different research groups' studies on *Bombina pachypus*[40], *Hyla sarda*[39], *Salamandra salamandra*[86], *Contracaecum osculatum*[87] sp. D as previously mentioned in this work.

De novo transcriptome assembly of *Bombina pachypus*

Six adult yellow-bellied toads have been examined, each displaying unique behavioural patterns: prolonged “unken-reflex” display versus no “unken-reflex” display designated as "+" and "-", respectively. Three toads demonstrated the prolonged “unken-reflex” (+), while the remaining three did not exhibit this reflex (−), as detailed in Table 1.

Bioproject	Run ID	Phenotypes	Raw sequences	Filtered sequences (% of the total reads)
PRJNA764013	SRR15927734	not unken-refex (-)	52’476’660	49’488’968 (94.31%)
PRJNA764013	SRR15927733	not unken-refex (-)	52’548’719	49’298’011 (93.81%)
PRJNA764013	SRR15927731	unken-refex (+)	53’007’498	49’675’835 (93.71%)
PRJNA764013	SRR15927732	not unken-refex (-)	54’267’911	50’813’920 (93.64%)
PRJNA764013	SRR15927730	unken-refex (+)	51’029’133	48’227’054 (94.51%)
PRJNA764013	SRR15927729	unken-refex (+)	52’999’652	49’850’617 (94.06%)
			316,329,573	297354405

Table 41 - A summary of the six libraries, which have been stored in the NCBI's Sequence Read Archive (SRA), detailing the count of both raw and trimmed reads for each sample.

A total of 316,329,573 raw reads were generated via *Illumina* sequencing and all were utilized for the assembly of the de novo transcriptome, as described in Chapter 3. As depicted in Figure 5, each read underwent a cleaning analysis process. Quality analysis was performed to estimate the RNA-Seq quality profiles, resulting in quality estimators for both raw and trimmed data. The quality assessment metrics for the trimmed data were subsequently aggregated across all samples into a single report for summary visualization, as illustrated in Fig. 9.

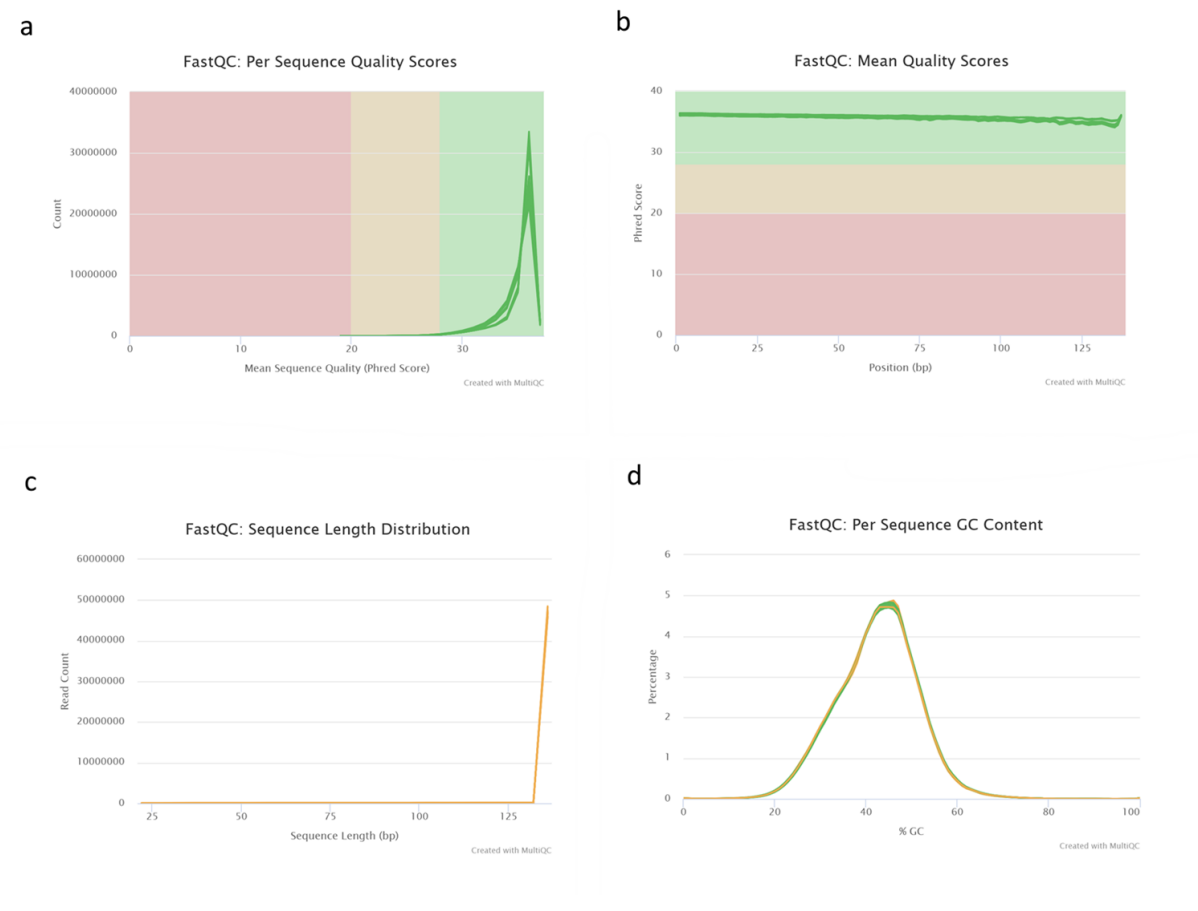


Figure 9 - The purified reads from each sample underwent evaluation using FastQC and were subsequently displayed using MultiQC. This included (a) the distribution of read counts in relation to average sequence quality, (b) the distribution of average quality scores, (c) the variation in read lengths, and (d) the GC content for each sequence.

After this cleaning step and the removal of low-quality reads, a total of 297,354,405 clean reads were retained, representing 94% of the original raw reads. These were subsequently used for de novo transcriptome assembly, as detailed in Table 1.

To construct an optimized de novo transcriptome and avoid chimeric transcripts, we employed rnaSPAdes, which automatically detected two k-mer sizes, approximately one-third and one-half of the maximal read length (the two detected k-mer sizes were 45 and 67 nucleotides, respectively). At this stage, a total of 1,118,671 assembled transcripts were generated by rnaSPAdes runs, with an average length of 689.41 bp and an N50 of 1474 bp, as reported in Table 2. The results from the assembly procedures were validated through three independent validator algorithms implemented in BUSCO, DETONATE, and TransRate. These tools generated several metrics used as a guide to evaluate error sources in the assembly process and to provide evidence about the quality of the assembled transcriptome.

The main metrics resulting from the assembly validators are shown in Table 2 (in the "After rnaSPAdes" column).

Validation scores	After rnaSPAdes	After CD-HIT-est	After Corset (Unigenes)
Basic parameters			
Total transcripts	1118671	896992	267959
Mean transcripts length (bp)	689.41	616.32	799
N50	1474	1082	2314
GC content (%)	40.15	39.93	39.87
Transrate v.1.0.3			
Transrate Assembly Score	0.0563	0.1281	—
Transrate Optimal Score	0.0877	0.178	—
Transrate Optimal Cutof	0.0138	0.0138	—
good contigs	991122	822552	—
p good contigs	0.89	0.92	—
Busco v.3.0.2			
Complete Buscos (C)	245 (96.1%)	240 (94.1%)	246 (96.5%)
Complete and single-copy Buscos (S)	157 (61.1%)	186 (72.9%)	157 (61.6%)
Complete and duplicated Buscos (D)	88 (34.5%)	54 (21.2%)	89 (25.3%)
Fragmented Buscos (F)	4 (1.6%)	8 (3.1%)	8 (3.1%)
Missing Buscos (M)	6 (2.3%)	7 (2.8%)	1 (0.4%)
Total Busco groups searched	255	255	255
Detonate v.1.9			
Score	−28610853057.11	−31067445817.60	—
BIC_penalty	−10912136.97	−11301742.60	—
Prior_score_on_contig_lengths_(f_function_canceled)	−2403141.71	−2100226.61	—
Prior_score_on_contig_sequences	−1072761315.55	−993903654.87	—
Data_likelihood_in_log_space_without_correction	−27529522562.34	−30064989468.40	—
Correction_term_(f_function_canceled)	−4746099.45	−4849274.88	—

Table 2 - Comparison rate of newly constructed transcripts against the de novo transcriptome of B. pachypus.

To ensure comprehensive data quality, we conducted evaluations using FastQC and MultiQC on all samples, both pre and post adaptor/sequence trimming. Our analysis indicated that the average read counts per quality score exceeded 35 (as seen in Fig. 10a). Similarly, average quality scores surpassed 35 for each base position (refer to Fig. 11b). The average lengths of sequences ranged between 126 and 130 bp (depicted in Fig. 11c), and the average GC content per sequence was approximately 40% (illustrated in Fig. 11d). For transcriptome assembly verification, we employed Busco, Detonate, and TransRate. The outcomes of this threefold validation process, detailed in Table 2, comprise scores from the trio of analytical tools, applied both before and after utilizing CD-HIT-est. By employing HISAT2, we confirmed that over 91% of the reads mapped back to the *B. pachypus* transcriptome assembly, signifying effective quality sequence reconstruction. Figure 10 delineates the count of raw reads, paired-reads post-trimming, and trimmed paired-reads mapped against the B. pachypus de novo transcriptome.

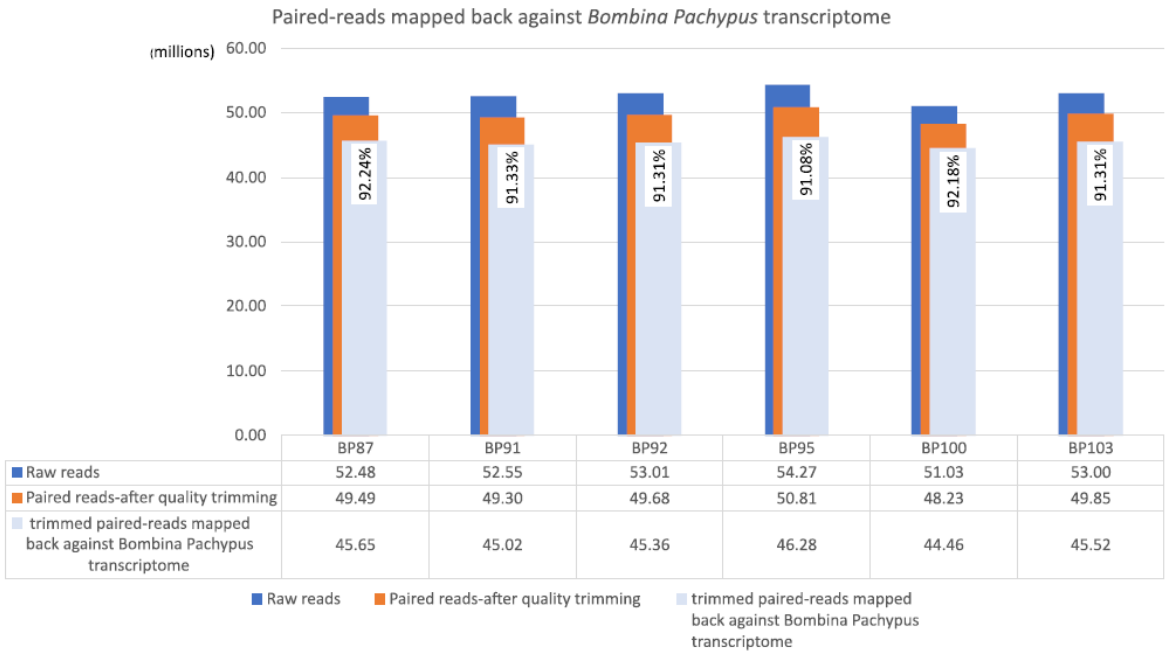


Figure 10 - For each sample, the total paired-reads are depicted in blue, while the total paired-reads post adapter removal and quality trimming are shown in orange. The trimmed paired-reads that have been aligned back to the B. pachypus assembled de novo transcriptome are represented in azure.

Notably, as shown in Table 2, the TransRate assessment revealed an increase in the "Transrate Optimal Score" following hierarchical clustering via CD-HIT-est, rising from 0.088 to 0.178. Likewise, the "Transrate Assembly Score" escalated from 0.056 to 0.128, more than doubling. The "good contigs" metric also saw an uptick post CD-HIT-est, attributable to redundancy elimination, with a final assembly score of 0.92. Each validation step reported a consistent GC value around 40%. Post CD-HIT-est, the transcriptome comprised 896,992 transcripts, with an average transcript length of 616.32 bp and an N50 of 1082 bp, demonstrating over 94% completeness as per Busco's assessment.

A significant advancement in redundancy reduction was achieved through the subsequent CORSET clustering step. The ultimate transcriptome assembly encompassed 267,959 transcripts, with an average length of 799 bp, an N50 of 2314, and a Busco assessment score exceeding 96%, thereby surpassing earlier results from the CD-HIT-est tool.

The results from all validation steps are detailed in Table 2. Finally, the CORSET output was processed with TransDecoder, the current standard tool for identifying long open reading frames (ORFs) in assembled transcripts, using the default settings.

For our de novo assembly process, we employed various annotation methods. Using DIAMOND, contigs were aligned against the Nr, SwissProt, and TrEMBL[96] databases to

retrieve the best annotations. An annotation matrix was created by selecting the top hit from each database. The analysis with BLASTX against Nr, SwissProt, and TrEMBL yielded 123,086 (64.57%), 77,736 (40.78%), and 122,907 (64.48%) contigs, respectively. In comparison, the BLASTP analysis resulted in 96,321 (50.53%), 57,877 (30.36%), and 97,256 (51.02%) contigs. Details of these datasets are summarized in Table 3.

Busco category	Nematoda	Metazoa	Eukaryota
Complete Buscos (C)	2676 (85,47%	793 (83,12%)	243 (95,30%)
Complete and single-copy Buscos (S)	1703 (54,40%	603 (63,10%)	192 (75,30%)
Complete and duplicated Buscos (D)	970 (31,00%)	190 (20,00%)	51 (20,00%)
Fragmented Buscos (F)	103 (3,30%)	33,39 (3,50%)	8
Missing Buscos (M)	354 (11,30%)	128 (13,40%)	4 (1,60%)
Total Busco groups searched	3131	954	255

Table 3 - Summary of homology annotation hits on the different databases queried in this study.

The BLASTX annotation output included a total of 77,391 sequences mapped across the Nr, SwissProt, and TrEMBL databases. The BLASTP annotation, on the other hand, resulted in 57,704 sequences mapped to these same databases. Venn diagrams in Fig. 43 illustrate the overlap of annotations across these databases for both DIAMOND BLASTX (Fig. 12a) and DIAMOND BLASTP (Fig. 12b)

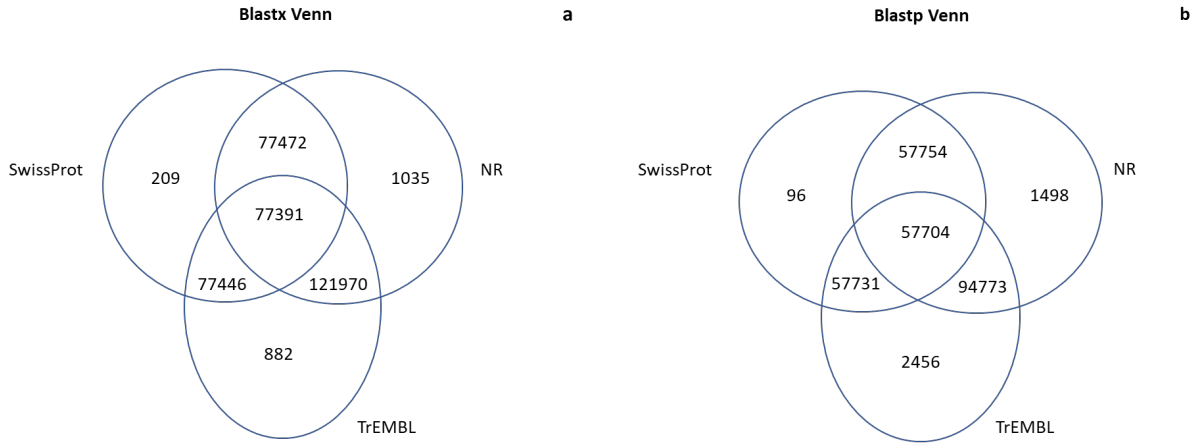


Figure 12 - Venn diagrams for the number of contigs annotated with DIAMOND (BLASTX (a) and BLASTP (b) functions) against the three databases: Nr, SwissProt, TREMBL.

InterProScan provided as result the corresponding InterPro accession numbers and, among other accession IDs, the GO and Kegg annotation. It produced a total of 32142 annotated contigs, 4747 contigs GO-annotated and 1025 contigs KEGG-annotated.

Additionally, Figs. 13 and 14 showcase the top ten species and gene products identified using BLASTX and BLASTP, respectively, by mapping the transcripts to the Nr reference database. BLASTX, which translates nucleotide sequences into protein sequences, provides more exhaustive results compared to BLASTP.

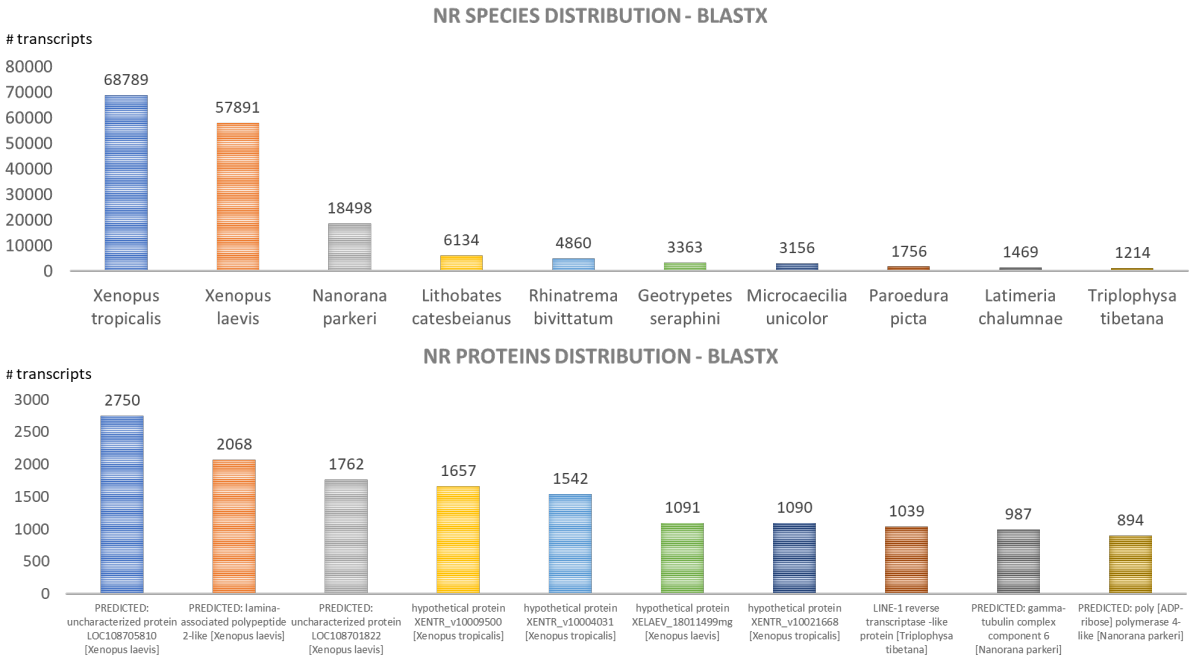


Figure 13 - Most represented species and gene product hits. Top 10 best species (a) and protein (b) hits present in the reference database (Nr, BLASTX).

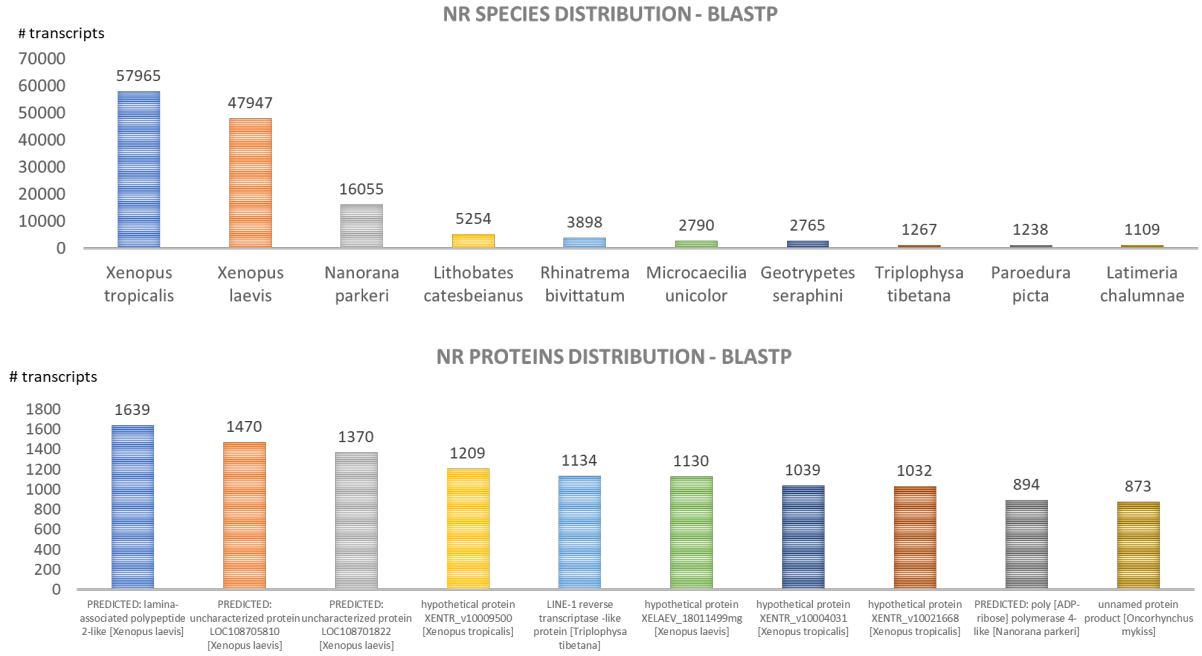


Figure 14 - Most represented species and gene product hits. Top 10 best species (a) and protein (b) hits present in the reference database (Nr, BLASTP).

Comparison with *Bombina orientalis* brain transcriptome

A comparative analysis of the brain de novo transcriptomes of *B. pachypus* and *B. orientalis* was performed[40]. The *B. orientalis* transcriptome, recently developed for a prey-catching conditioning study, is available in the GEO archive of NCBI at this link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171766. To align the datasets for comparison, we initially applied Transdecoder to predict open reading frames (ORFs) in the *B. orientali*s transcriptome, adhering to the default parameters.

The findings of the B. orientalis ORF prediction are accessible on Figshare at https://doi.org/10.6084/m9.fgshare.20319633/. We also employed DIAMOND's makedb function to create an indexed protein database. The next step involved aligning B. pachypus's predicted coding sequences and proteins (the query files) against the B. orientalis protein database (the reference) via DIAMOND BLASTX and BLASTP, which resulted in 167,041 and 156,248 matches, respectively.

De novo transcriptome assembly of *Hyla sarda*

For assembling a comprehensive de novo transcriptome that captures the geographical and behavioural diversity, nine Tyrrhenian tree frog individuals from different islands were selected and sequenced: four from Sardinia, four from Corsica, and one from Elba, as detailed in Table 4. To determine their levels of boldness and shyness, two distinct tests were employed. The first test examined the frogs' willingness to explore new environments, while the second measured their delay in emerging from a shelter, indicative of their risk-taking behaviour. Detailed descriptions of the sampling process and these behavioural tests can be found in Bisconti et al. (2022)[99]. The specific behavioural traits of each frog are documented in Table 4.

Sample code	Sample location	Phenotypes	Run ID	Raw sequences	Filtered sequences (% of the total reads)
HS 9	Porto Pollo, Sardinia	Bold	ERR9486926	72098178	69503761 (96.4)
HS 17	Porto Pollo, Sardinia	Bold	ERR9486869	53786371	52007828 (96.7)
HS 27	Cala Ginepro, Sardinia	Bold	ERR9486870	54282464	52269332 (96.3)
HS 44	Cala Ginepro, Sardinia	Bold	ERR9486863	49300066	47606858 (96.6)
HS 56	Elba	Bold	ERR9433479	51227766	46626281 (91.0)
HS 87	Etang de Canettu, Corsica	Shy	ERR9433456	53775786	50544832 (94.0)
HS 93	Etang de Canettu, Corsica	Shy	ERR9433396	54008324	50503363 (93.5)
HS 104	Aleria, Corsica	Shy	ERR9433394	51829828	47628168 (91.9)
HS 111	Aleria, Corsica	Shy	ERR9433393	52754498	49148365 (93.2)

Table 4 - Compilation of the nine samples along with their corresponding data stored in the ENA database (European Nucleotide Archive), including the Study Accession ID details.

For initiating the de novo transcriptome assembly, high-quality RNA-Seq sequence reads were utilized as the primary input. The initial stage of processing raw data involved the elimination of adapters and low-quality sequences, resulting in 456,838,788 clean reads, which accounted for 93% of the initial raw reads. These clean reads were then retained for constructing the de novo transcriptome assembly. A comprehensive summary of the sequencing data, including the total count of reads both before and after the trimming process, is provided in Table 4.

Validation scores	Trinity	rnaSPAdes	CD-HIT-est (unigenes)
Transrate v.1.0.3			
Total transcripts	1580186	1244846	1295741
N50	684	1475	914
GC Content (%)	44.00	44.00	44.00
Transrate Assembly Score	0.0277	0.0288	0.0519
Transrate Optimal Score	0.0936	0.0459	0.0933
Transrate Optimal Cutof	0.3149	0.0155	0.0144
good contigs	0.48	0.88	0.85
Busco v.3.0.2			
Complete Buscos (C)	253 (99.2 %)	247 (96.8 %)	251 (98.4 %)
Complete and single-copy Buscos (S)	80 (31.4 %)	97 (38.0 %)	124 (48.6 %)
Complete and duplicated Buscos (D)	173 (67.8 %)	150 (38.8 %)	127 (49.8 %)
Fragmented Buscos (F)	1 (0.4 %)	5 (2.0%)	2 (0.8 %)
Missing Buscos (M)	1 (0.4 %)	3 (1.2%)	2 (0.8 %)
Total Busco groups searched	255	255	255

Table 5 - Information detailing the outputs derived from rnaSPAdes and Trinity, coupled with the results post the application of CD-HIT-est (Unigenes), with a specific focus on the Trinity-related data.

Subsequently, the quality metrics for the trimmed data were compiled from all samples into a consolidated report. This report provided a summarized visual representation of the data quality, as depicted in Figure 15.

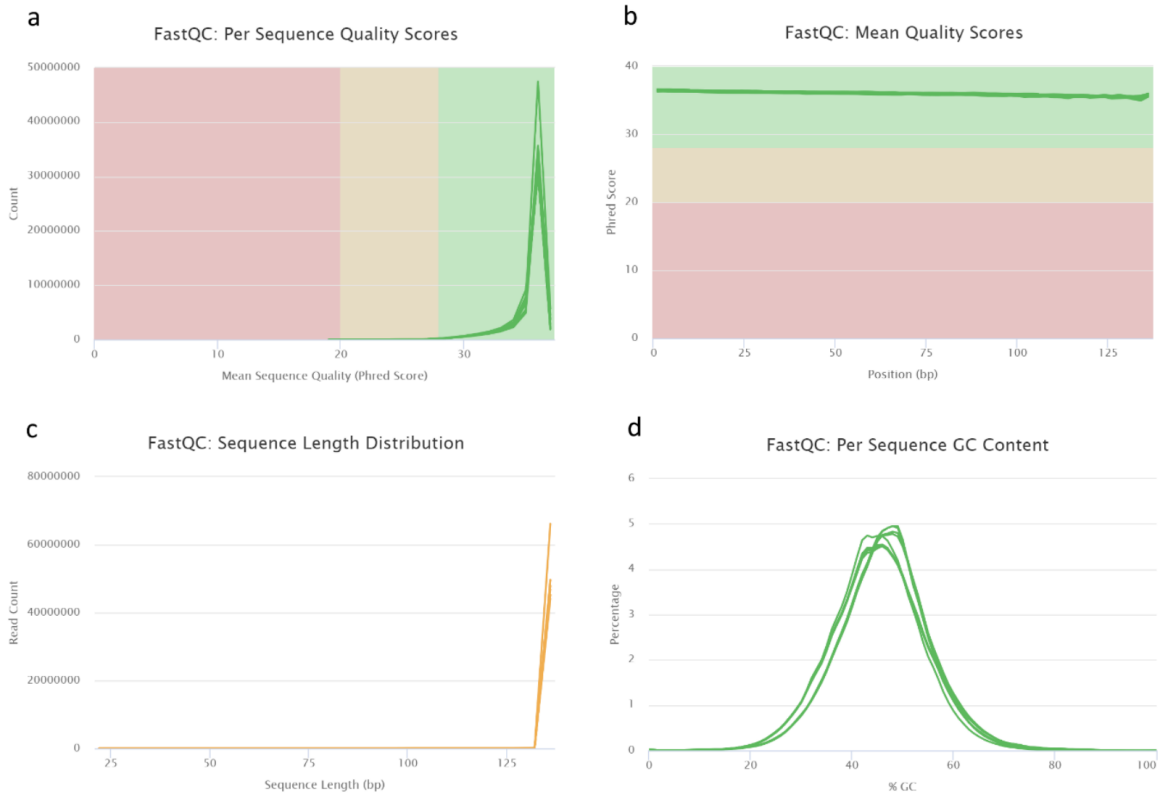


Figure 15 - Analysis of cleaned reads from all samples conducted with FastQC and presented through MultiQC visualization. The components include (a) distribution of read counts by average sequence quality, (b) distribution of average quality scores, (c) distribution of read lengths, and (d) distribution of mean quality scores.

To determine the most effective strategy for transcriptome assembly, two different workflows were evaluated: Trinity and rnaSPAdes. Both assemblers were employed with the aim of enhancing the final assembly's reliability. The outcomes derived from using each of these assembly methods are presented in Table 6.

Validation scores	Trinity	rnaSPAdes	CD-HIT-est (Unigenes)
Transrate v.1.0.3			
Total transcripts	1580186	1244846	1295741
N50	684	1475	914
GC content (%)	44.00	44.00	44.00
TransRate Assembly Score	0.0277	0.0288	0.0519
TransRate Optimal Score	0.0936	0.0459	0.0933
TransRate Optimal Cutoff	0.3149	0.0155	0.0144
p good contigs	0.48	0.88	0.85
Busco v.3.0.2			
Complete Buscos (C)	253 (99.2%)	247 (96.8 %)	251 (98.4%)
Complete and single-copy Buscos (S)	80 (31.4 %)	97 (38.0 %)	124 (48.6%)
Complete and duplicated Buscos (D)	173 (67.8 %)	150 (38.8 %)	127 (49.8 %)
Fragmented Buscos (F)	1 (0.4 %)	5 (2.0 %)	2 (0.8 %)
Missing Buscos (M)	1 (0.4%)	3 (1.2 %)	2 (0.8 %)
Total Busco groups searched	255	255	255

Table 6 - Statistical data showcasing the output from rnaSPAdes and Trinity, and the subsequent results following the application of CD-HIT-est (Unigenes).

The output from both Trinity and rnaSPAdes assemblers was combined using Trans-AbySS, leading to the selection of the most effective contigs. The merged assembly was then used as the starting point for CD-HIT-est processing. The details of these results are outlined in Table 6. Following this, the CD-HIT-est output was utilized for predicting open reading frames (ORFs) as part of the transcriptome annotation process.

The TransRate evaluation revealed high scores for the "TransRate Optimal Score," consistently around 0.09. After applying CD-HIT-est, which led to the reduction of redundancy, there was a decrease in the "good contigs" score, settling at 0.85 in the final assembly. TransRate also indicated a GC content of approximately 44%. The ultimate transcriptome assembly comprised 1,295,741 transcripts with an N50 value of 914 bp.

Furthermore, the BUSCO assessment showed a notable score for the "Complete and single-copy" category, reaching 124 (48.6%) following the CD-HIT-est process. As indicated in Table 6, this step was effective in diminishing redundancy, reducing the number of transcripts generated by the two initial assemblers, and enhancing overall quality scores.

The process of mapping paired-reads back against the *Hyla sarda* de novo transcriptome is illustrated in Figure 6. This figure showcases the quantities of raw reads, paired reads post-quality trimming, and trimmed paired reads aligned with the *H. sarda* de novo transcriptome, with values consistently exceeding 90%, represented in millions.

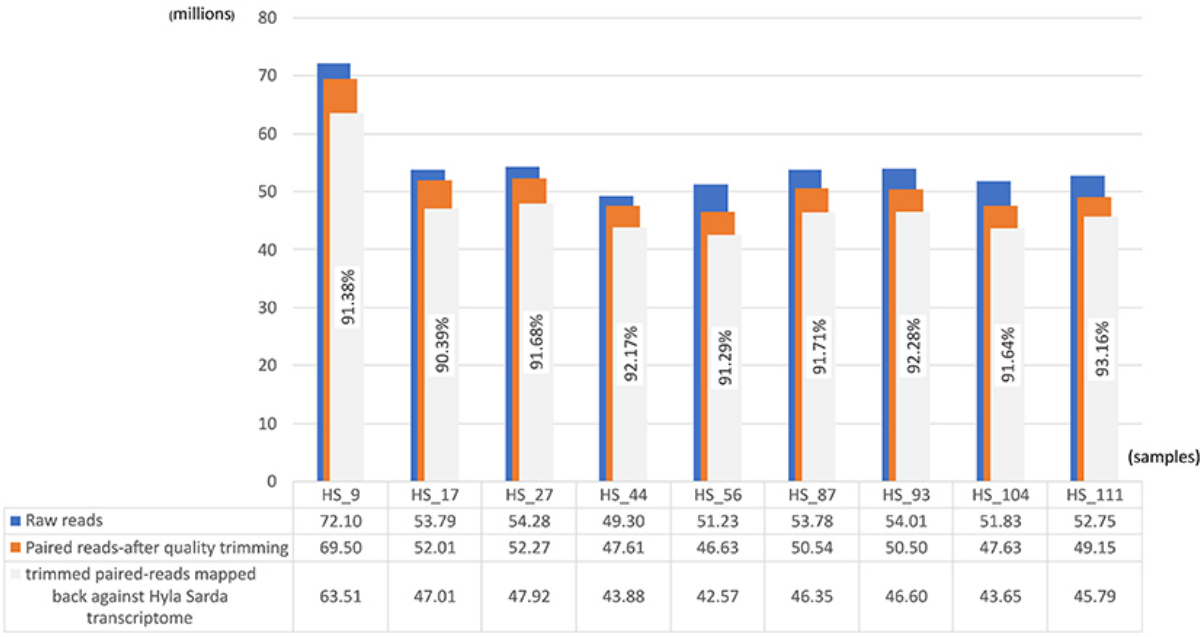


Figure 16 - Alignment of paired-reads with the Hyla sarda de novo transcriptome. Each sample is represented with the total number of paired-reads in blue, the count of paired-reads post-adapter removal and quality trimming in orange, and the trimmed paired-reads realigned with the Hyla sarda de novo assembled transcriptome in gray.

In each sample, the total number of paired-reads is depicted in blue. The paired-reads remaining after adapter removal and quality trimming are shown in orange. The trimmed paired-reads that have been realigned with the *Hyla sarda* de novo assembled transcriptome are represented in gray.

The subsequent analyses incorporated both homology and functional annotations. The Open Reading Frames (ORFs) that were predicted were aligned using DIAMOND, employing the BLASTX and BLASTP functions, against three databases: NCBI NR, Swiss-Prot, and TrEMBL.The results of these annotations are stored in six output files, which are accessible on Figshare and detailed in Table 7.

Label	Name of data file/dataset	File types	Data repository and identifier (DOI or accession number)
Data file 1	Trinity RNA-Seq de novo transcriptome assembly	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=3145760812
Data file 2	rnaSPADES RNA-Seq de novo transcriptome assembly	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457605
Data file 3	Hyla Sarda RNA-Seq de novo transcriptome assembly (Unigenes)	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457614
Data file 4	Open reading frames (ORFs) prediction	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457617
Data file 5	Homology annotation from nonredundant (nr) NCB	Fasta file (.fa)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=3680437234
Data file 6	Homology annotation from SwissProt	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804381
Data file 7	Homology annotation from TrEMBL UniProt	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804405

Label	Name of data file/dataset	File types	Data repository and identifier (DOI or accession number)
Data file 1	Trinity RNA-Seq de novo transcriptome assembly	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=3145760812
Data file 2	rnaSPADES RNA-Seq de novo transcriptome assembly	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457605
Data file 8	Homology annotation from nonredundant (nr) protein NCBI	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804348
Data file 9	Homology annotation from SwissProt Protein	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804357
Data file 10	Homology annotation from TrEMBL UniProt Protein	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804363
Data file 11	Detailed full Homology annotation Best Hits	Xlsx file (.xlsx)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=368102375
Data file 12	InterProScan results	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=368135224

Table 7 - A consolidated view of the data files and datasets produced in this study, along with information regarding the repositories where they are archived.

Upon analyzing the BLASTX results against the NCBI Nr, Swiss-Prot, and TrEMBL databases, the respective outcomes were as follows: 179944 sequences (46.99%), 133455 (34.85%), and 179978 (47.00%). Similarly, the BLASTP analysis against the same databases – NCBI Nr, Swiss-Prot, and TrEMBL – yielded 128401 sequences (33.53%), 88996 (23.24%), and 130165 (33.99%). Detailed data on the homology annotation for each of these three databases can be found in the Supplementary Table 8.

Annotation statistics	
Number of BLASTX results	
NCBI nr	179,944 (46.99 %)
Swiss-Prot	133,455 (34.85 %)
TrEMBL	179,978 (47.00 %)
Number of BLASTP results	
NCBI nr	128,401 (33.53 %)
Swiss-Prot	88,996 (23.24 %)
TrEMBL	130,165 (33.99 %)

Table 8 - A comprehensive depiction of annotations distributed among different databases.

A Venn diagram was developed to illustrate the overlap in annotations across various databases. These diagrams were prepared for both Diamond BLASTX and Diamond BLASTP results, as depicted in Figure17A and Figure 17B, respectively. They demonstrate that 132635 unigenes were consistently annotated across the three databases in the BLASTX results, while 88671 unigenes were similarly annotated in the BLASTP results.

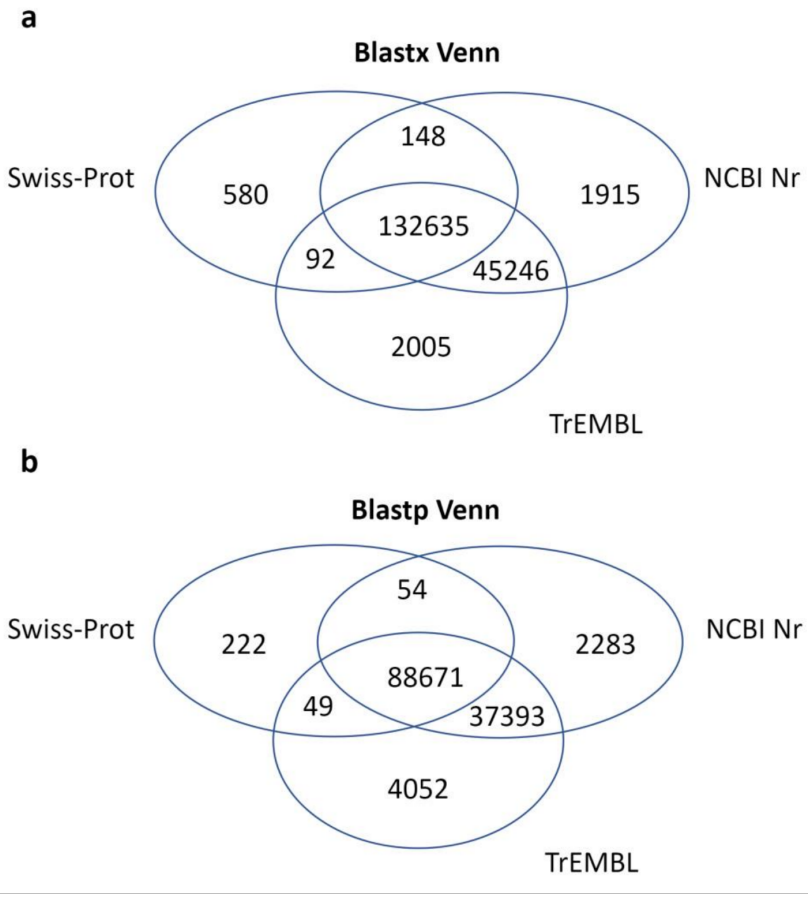


Figure 17 - Venn diagrams depicting the quantity of contigs annotated using Diamond, employing both Blastx (a) and Blastp (b) functions, across three databases: NCBI nr, Swiss-Prot, and TrEMBL.

For a comprehensive and current homology annotation file with detailed insights on the top hits, the Uniprot API was utilized to extract information about protein IDs that correspond to the predicted ORF in the TrEMBL database. The outcomes of these queries were compiled into a file in xlsx format, which is accessible on Figshare (Table 7, Data file 11, "Detailed full Homology annotation Best Hits.xlsx"). This file includes more extensive details than the Diamond output, encompassing aspects like database, locus name, sequence description, sequence length, division, gene name, organism, protein name, and product name.

Furthermore, the predicted ORF underwent a secondary processing step using InterProScan for scanning InterProScan signatures. The results from this InterProScan analysis are also available on Figshare (Table 7, Data file 12, "Interproscan Results.tsv").

De novo transcriptome assembly of *Salamandra salamandra*

Salamander larvae from a population have been collected in Central Italy showing behavioural polymorphism[86]. Details about sampling, housing and behavioural essays are described in Chiocchio et al. [24]. Ten larvae of fire salamander have been selected like representative of two distinct behavioural profiles, i.e., slow, less active, and less exploratory behaviour vs fast, more active and more exploratory behaviour (thereafter referred as "slow" and "fast", respectively; see Table 9).

Sample	Phenotype	Run ID	Raw sequences	Filtered sequences	% Trimmed reads
R530	slow	ERR8963532	148,383,674	143,289,768	96.57
R531	fast	ERR8971203	113,170,212	108,848,306	96.18
R538	slow	ERR8971371	131,207,312	126,584,688	96.48
R541	fast	ERR8962162	10,147,648	97,683,726	96.27
R547	fast	ERR8971605	116,726,758	112,808,264	96.64
R560	Slow	ERR8971694	123,732,768	119,419,898	96.51
R564	Slow	ERR8971822	99,188,392	95,748,700	96.53
R565	fast	ERR8971822	98,856,326	95,578,524	96.68
R572	fast	ERR8972978	98,856,326	113,636,400	96.43
R573	slow	ERR8974269	102,857,670	98,754,492	96.02

Table 9 - An overview of 10 libraries stored in the ENA (European Nucleotide Archive) under Study Accession Id PRJEB51202, detailing the count of both raw and trimmed reads for each sample.

A dataset comprising 1,153,432,918 paired reads was utilized for the assembly of the de novo transcriptome. The raw reads underwent quality trimming to eliminate low-quality bases and adapter sequences, with unpaired reads being discarded. Following this cleaning step, a total of 1,112,352,766 high-quality reads (representing 96% of the raw reads) were retained for the construction of the de novo transcriptome assembly, as shown in Table 9.

The quality of each individual raw read was assessed using FastQC. The results from all samples were then aggregated into a comprehensive report using MultiQC, as depicted in Figure 18.

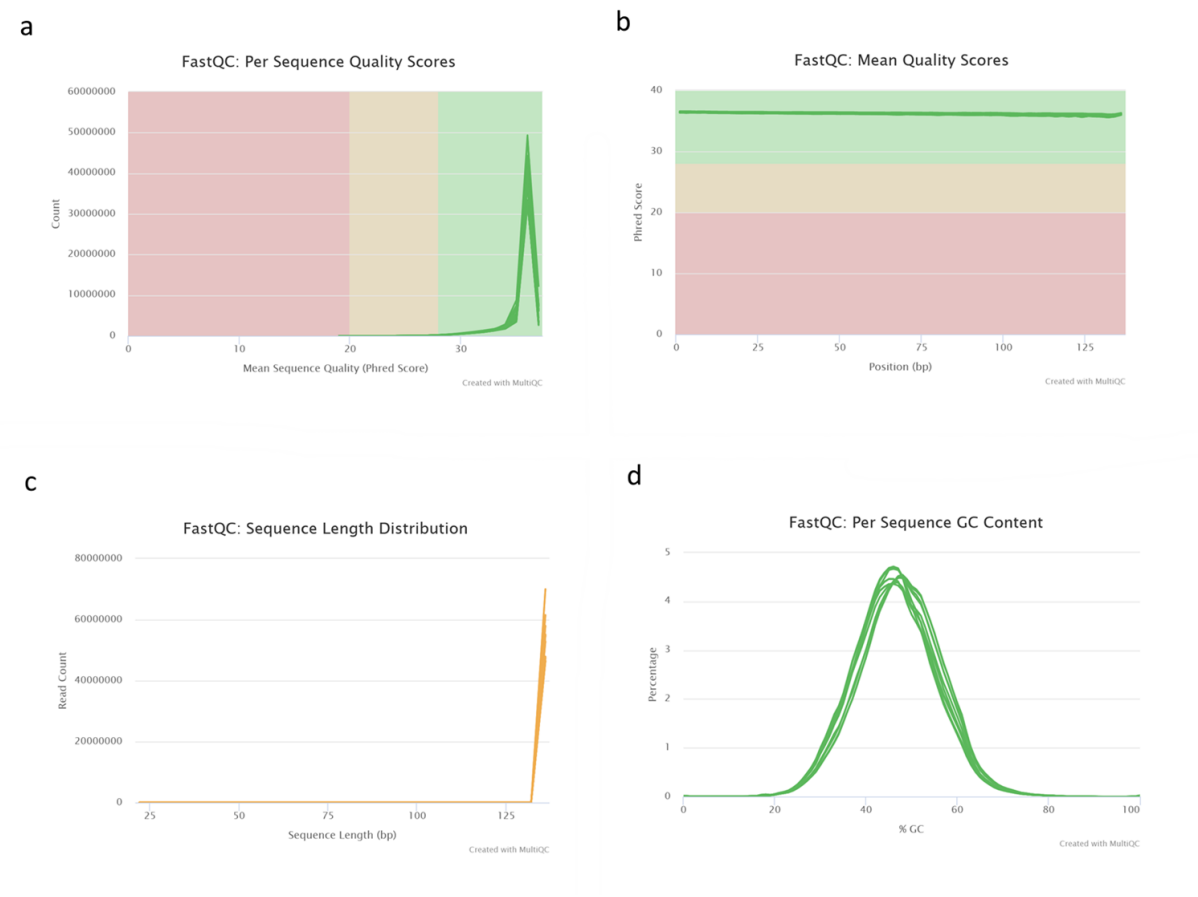


Figure 18 - Evaluation of cleaned reads from all samples performed using FastQC, with results displayed via MultiQC. The analysis encompasses (a) the distribution pattern of read counts based on average sequence quality, (b) the spread of average quality scores, (c) the variation in read lengths, and (d) the distribution of mean quality scores.

A de novo transcriptome assembly was conducted using both the Trinity and rnaSPAdes assemblers. Trinity generated a total of 1,207,872 transcripts, while rnaSPAdes produced 1,094,271 transcripts, as detailed in Table 10.

Validation scores	Trinity	rnaSPAdes	CD-HIT-est (Unigenes)
Basic parameters			
Total transcripts	1,207,872	1,094,271	1,146,571
N50	1742	1979	1529
GC content (%)	45.0	45.0	44.7
Transrate v.1.0.3			
TransRate Assembly Score	0.05	0.06	0.06
TransRate Optimal Score	.1	0.09	0.1
TransRate Optimal Cutoff	0.01	0.01	0.01
good contigs	947,064	974,586	972,431
p good contigs	0.8	0.9	0.9
DETONATE v.1.9			
Score	−4.8e10	−3.9e10	−4.3e10
BIC_penalty	−1.2e7	−1.1e7	−1.1e7
Prior_score_on_contig_lengths_(f_function_canceled)	−2.5e6	−2.3e6	−2.1e6
Prior_score_on_contig_sequences	−1.3e9	−1.1e9	−1.1e9
Data_likelihood_in_log_space_without_correction	−4.6e10	−3.8e10	−4.2e10
Correction_term_(f_function_canceled)	−5.0e6	−4.9e	−5.1e6

Table 10 - Statistics on rnaSPAdes and Trinity output and the result after CD-HIT evaluated with the two assembly validators.

After completing the assembly process, the outputs from both Trinity and rnaSPAdes were combined using Trans-AbySS, employing its merging function. This merged output then served as the input for subsequent stages.

The assembly results underwent two stages of validation: an initial evaluation of the preliminary assembly and a subsequent assessment of the final, non-redundant assembly output. Both TransRate and DETONATE were utilized to generate a variety of metrics, which helped in identifying potential errors in the assembly process and in assessing the quality of the assembled transcriptome. Table 10 presents the results of these assessment analyses, including a) the assembly output from Trinity, b) the assembly output from rnaSPAdes, and c) the final assembly output, which is the merged assembly with removed redundancies.

Final assembly has been furthermore evaluated analyzing the gene content by launching BUSCO, on four databases of ortholog genes: CVG (Core Vertebrate Genes), Tetrapoda, Vertebrata and Eukariota databases. In Table 11 we reported BUSCO results.

Busco category	Tetrapoda Database	Vertebrata Database	CVG (Core Vertebrata Genes)	Eukaryota Database
Complete Buscos (C)	3723 (94.2%)	2456.7 (95.3%)	226 (97.4%)	255 (98.4%)
Complete and single-copy Buscos (S)	1975 (50.0%)	1318.9 (50.7%)	111.84 (48.5%)	96.9 (37.6%)
Complete and duplicated Buscos (D)	1739 (44.2%)	1163.7 (44.6%)	114.2 (48.9%)	155.5 (60.8%)
Fragmented Buscos (F)	119 (3.3%)	103.4 (3.5%)	4.6 (2.1%)	5.1 (1.6%)
Missing Buscos (M)	118.5 (2.5%)	25.9 (1.2%)	2 (0.5%)	0.0%

Total Busco groups searched	950	2586	233	255

Table 11 - presents the results of the BUSCO (v. 5) validation carried out through the gVolante web server, which encompassed four distinct databases: Tetrapoda, Vertebrata, Core Vertebrate Genes (CVG), and Eukaryota.

Additionally, Fig. 19 illustrates the distribution of completed, fragmented, and missing genes as mapped from the four databases. Notably, a substantial percentage of genes were found to be complete in both the Tetrapoda and Vertebrata databases, affirming the high quality of our assembly.

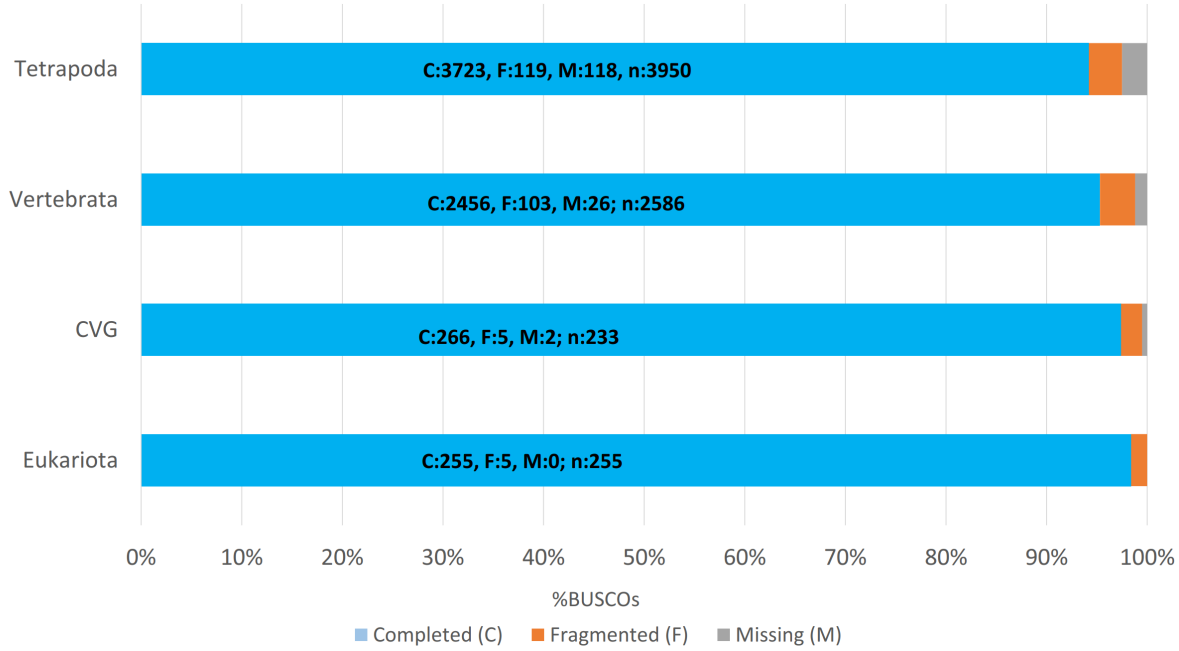
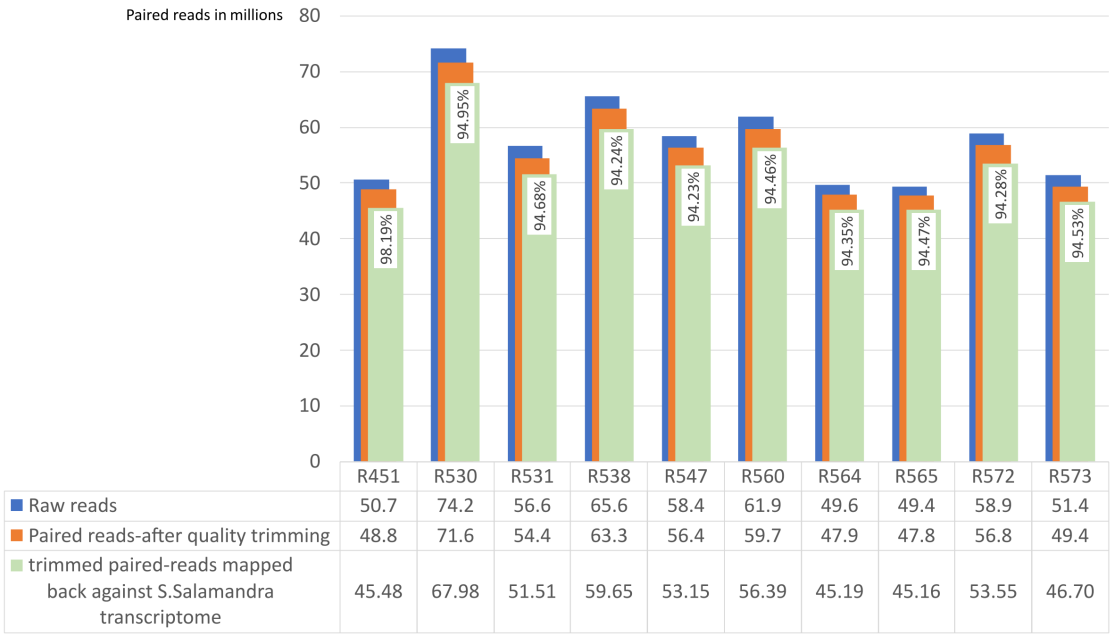


Figure 19 – BUSCO assessment results under a graphical point of view.

An additional method of quality assessment involved aligning the trimmed raw data with the final de novo transcriptome assembly, using HISAT2. The results of this mapping are displayed in Figure 20. Notably, the mapping percentages consistently exceeded 94%, further confirming the high quality of the assembly.



	R451	R530	R531	R538	R547	R560	R564	R565	R572	R573
Raw reads	50.7	74.2	56.6	65.6	58.4	61.9	49.6	49.4	58.9	51.4
Paired reads-after quality trimming	48.8	71.6	54.4	63.3	56.4	59.7	47.9	47.8	56.8	49.4
trimmed paired-reads mapped back against S.Salamandra transcriptome	45.48	67.98	51.51	59.65	53.15	56.39	45.19	45.16	53.55	46.70

Figure 20 – In each sample, the data is visually represented using a color-coded scheme: the total number of paired-reads is shown in blue, the paired-reads that remain after adapter removal and quality trimming are displayed in orange, and the trimmed paired-reads that have been successfully mapped back against the assembled de novo transcriptome of *Salamandra salamandra* are highlighted in azure.

Different annotations have been employed to perform further analysis obtaining the results shown in Table 12 obtained align contigs against DIAMOND on NCBI, Swiss-Prot and TrEMBL to retrieve the corresponding best annotations.

Annotation statistics	
Number of BLASTX results	
NCBI nr	220,041 (49.86 %)
Swiss-Prot	154,324 (34.97 %)
TrEMBL	220,521 (49.97 %)
Number of BLASTP results	
NCBI nr	152,278 (34.50 %)
Swiss-Prot	95,566 (21.88 %)
TrEMBL	154,341 (34.97 %)

Table 12 - Overview of annotations across various databases.

Overview of data files and data sets produced in this study, with information on data repository and accession numbers, are summarized in Table 13.

Label	Name of data file/dataset	File types	Data repository and identifier (DOI or accession number)
Data file 1	S. salamandra Trinity de novo transcriptome assembly	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22341469
Data file 2	S. salamandra rnaSPAdes de novo transcriptome assembly	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22341526
Data file 3	S. salamandra merged assembly	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22341439
Data file 4	S. salamandra unigenes	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22341541
Data file 5	S. salamandra Open Reading Frames (ORFs) prediction	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22341550
Data file 6	S. salamandra homology annotation (blastx), NR	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341577
Data file 7	S. salamandra homology annotation (blastx), Swiss-Prot	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341589
Data file 8	S. salamandra homology annotation (blastx), TrEMBL	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341673
Data file 9	S. salamandra homology annotation (blastp), NR	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341595
Data file 10	S. salamandra homology annotation (blastp), Swiss-Prot	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341679
Data file 11	S. salamandra homology annotation (blastp), TrEMBL	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341694
Data file 12	S. salamandra functional annotation InterProScan results	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22341715
Data file 13	P. waltl transcriptome	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22341739
Data file 14	P. waltl Open Reading Frames (ORFs) prediction	Fasta file (.fa)	https://doi.org/10.6084/m9.figshare.22680325
Data file 15	S. salamandra CDSs vs P. waltl ORFs (by blastx)	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22680244
Data file 16	P. waltl CDSs vs S. salamandra ORFs (by blastx)	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22680259
Image file 1	S. salamandra MultiQC quality assessment	PDF file (.pdf)	https://doi.org/10.6084/m9.figshare.22718221

Table 13 - Summary of data files and datasets generated in this research, including details about the data repository where they are stored.

A Venn diagram was created to show the redundancy of the annotations in different databases; the diagrams were constructed for both DIAMOND BLASTX and DIAMOND BLASTP (Figure 21) and showed 153,048 (BLASTX) and 95,942 (BLASTP) shared unigenes, i.e., annotated from the three databases.

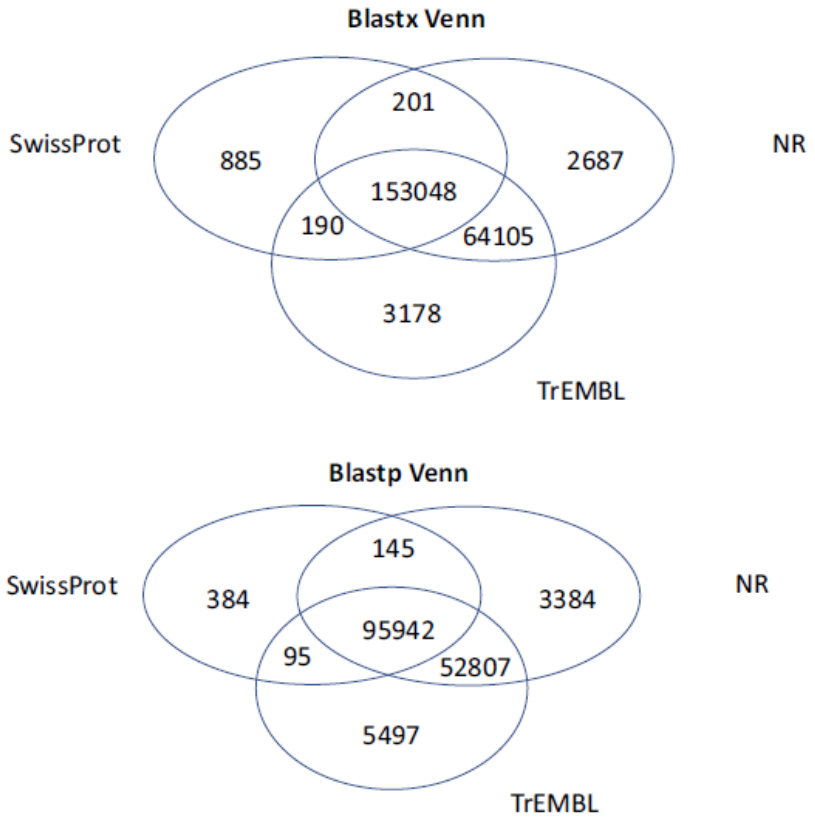
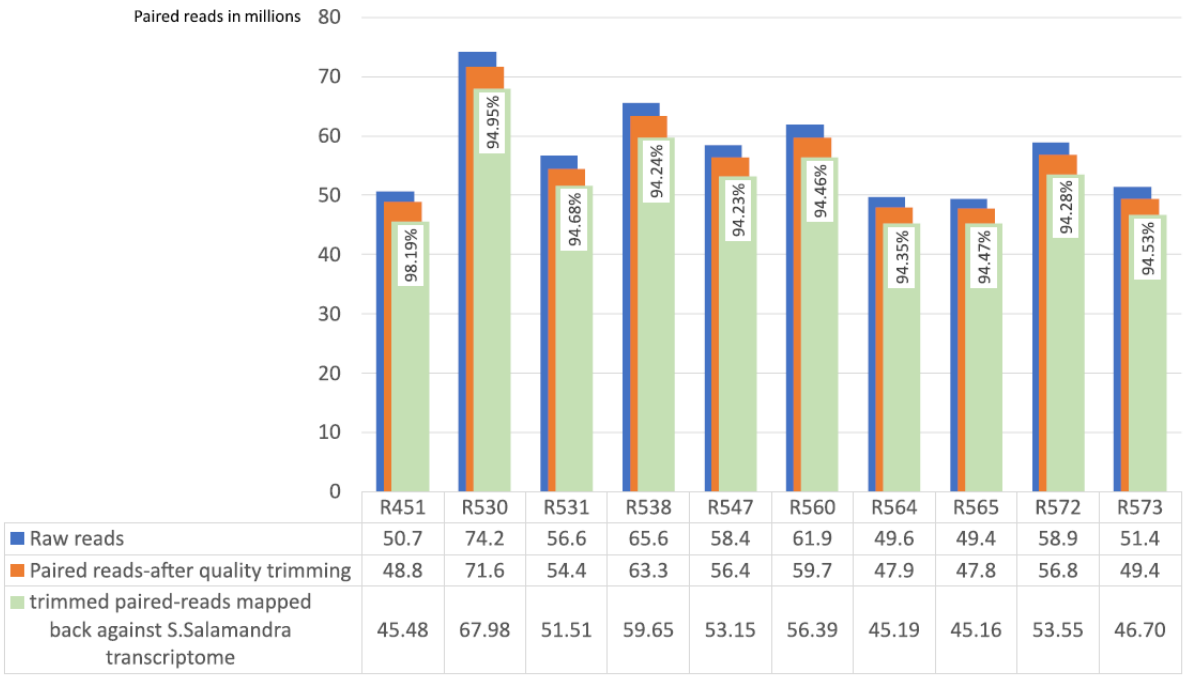


Figure 21 - Venn diagrams depicting the count of contigs annotated using DIAMOND (through both BLASTX and BLASTP functions), across three databases: NCBI nr, Swiss-Prot, and TREMBL.

Subsequently, the contigs underwent processing with InterProScan for protein signature prediction, resulting in 56,179 annotated unigenes. Of these, 9,850 were GO-annotated and 2,311 were KEGG-annotated. The final validation involved aligning the trimmed reads with the de novo assembled transcriptome of *S. salamandra* using HISAT2. This step revealed an impressive mapping percentage of over 94% (as seen in Figure 22), underscoring the high quality of the assembly. After applying CD-HIT-est, the ultimate transcriptome (unigenes) comprised 1,146,571 transcripts with an N50 of 1529 bp, exhibiting more than 94% completeness in the BUSCO evaluation across each of the queried databases.



	R451	R530	R531	R538	R547	R560	R564	R565	R572	R573
Raw reads	50.7	74.2	56.6	65.6	58.4	61.9	49.6	49.4	58.9	51.4
Paired reads-after quality trimming	48.8	71.6	54.4	63.3	56.4	59.7	47.9	47.8	56.8	49.4
trimmed paired-reads mapped back against S.Salamandra transcriptome	45.48	67.98	51.51	59.65	53.15	56.39	45.19	45.16	53.55	46.70

Figure 22 - Each sample is visually represented with total paired-reads in blue, while the paired-reads post adapter removal and quality trimming are shown in orange. The trimmed paired-reads that have been aligned back to the Salamandra salamandra assembled de novo transcriptome are indicated in azure.

Comparison with *Pleurodeles waltl*

We utilized the reference genome and its corresponding genome annotation of *Pleurodeles waltl* (the P. waltl gene model) to achieve three objectives: (1) extract the P. waltl transcriptome, (2) predict its ORFs (Open Reading Frames), and (3) conduct a reciprocal mapping of the predicted ORFs between *S. salamandra* and *P. waltl*. This approach aimed to evaluate the similarity rate between our de novo transcriptome of *S. Salamandra* and the *P. waltl* transcriptome.

Initially, we downloaded the genome and annotation files of *P. waltl*. Using GffRead[98], a tool for manipulating GFF and GTF format files, we extracted the P. waltl transcriptome. The annotation file (aPleWal.anno.v2.20220926.gff3) and the assembled genome (aPleWall.pri.20220803.fasta.gz, approximately 20 GB) served as input for GffRead. The resulting transcriptome file, P_Waltl_transcripts.fa, was subsequently uploaded to figshare.

Next, we used TransDecoder to predict the ORFs of P. waltl, creating a BLAST database from these. We then applied DIAMOND (using the blastx function) to compare S. salamandra's ORFs (query sequences in multi-fasta format) against P. waltl's indexed ORF blast database. Out of 441,339 CDS-ORF sequences in S. salamandra, 290,095 (65.7%) matched the ORFs of P. waltl. Conversely, when mapping P. waltl's CDS-ORFs (1,180,470)

against S. Salamandra's ORFs, 792,563 (67.1%) showed similarity. The output files from both DIAMOND runs, salamandra_unigenes_vs_pleurodeles_orf_blastx.tsv and pleurodeles_vs_salamandra_unigenes_orf_blastx.tsv, were also uploaded to figshare.

This methodology provided a comparative analysis of the ORFs from S. salamandra and P. waltl, illustrating the genomic similarities between these two species.

De novo transcriptome assembly of *Contracaecum osculatum*

Contracaecum osculatum (sensu lato) third-stage larvae (L3) were extracted from the body cavity of the ice fish, *Chionodraco hamatus* caught in the Ross Sea, Antarctica[87].

The RNA-Seq data collected on *C. osculatum sp. D* L3 underwent processing for extensive transcriptome sequencing. The methodology of the bioinformatics processes, derived from methodologies in three prior studies, is depicted in Figure 5 in Chapter 3. The process generated a total of 425,954,724 read pairs. Each of these pairs was subjected to a cleaning and analysis phase using FastQC both prior to and following the trimming stage.

The initial raw reads were subjected to a quality trimming process using Trimmomatic excluding any unpaired read from subsequent steps in the assembly pipeline. Following this cleaning phase and the removal of low-quality reads, a total of 387,489,135 reads remained. These retained reads represent 92% of the original raw reads, as detailed in Table 14.

Run ID	Temperature	Raw sequences	Filtered sequences	% Trimmed reads
SRR23456143	−2 °C	22,634,131	20,736,790	91,62
SRR23456144	−2 °C	21,025,371	18,647,738	88,69
SRR23456145	−2 °C	24,914,886	22,018,652	88,38
SRR23456146	−2 °C	25,226,687	23,011,308	91,22
SRR23456147	−2 °C	25,192,198	23,000,985	91,30
SRR23456157	−2 °C	23,345,551	21,297,927	91,23
SRR23456158	−2 °C	23,026,402	20,808,470	90,37
SRR23456148	1 °C	22,162,833	23,000,985	89,26
SRR23456149	1 °C	22,937,773	20,536,345	89,53
SRR23456150	1 °C	20,049,583	18,077,263	90,16
SRR23456159	1 °C	22,387,210	20,237,154	90,40
SRR23456160	1 °C	20,747,424	18,508,542	89,21
SRR23456156	37 °C	33,435,950	29,964,089	89,62
SRR23456155	37 °C	20,858,439	19,085,468	91,50
SRR23456154	37 °C	24,951,744	22,987,951	92,13
SRR23456153	37 °C	25,963,075	23,158,637	89,20
SRR23456152	37 °C	24,076,721	21,753,688	90,35
SRR23456151	37 °C	23,018,746	20,657,143	89,74

Table 14 - Overview of the 18 Libraries Submitted to the NCBI Sequence Read Archive (SRA) under Project SRP42248343. Details for Each Sample (from Three Whole L3 Larvae) Include Run ID, Exposure Temperature, and Count of Raw, Filtered, and Trimmed Reads.

Due to the lack of a reference genome for *C. osculatum* sp. D, a de novo transcriptome assembly process was performed using rnaSPAdes that automatically identified two k-mer sizes that were approximately one-third and half the length of the longest read, specifically 45 and 67 nucleotides. This process resulted in the creation of 237,314 assembled contigs, with an N50 of 1783 bp, as detailed in Table 15.

Validation scores	rnaSPAdes output	Corset output (unigenes)
Basic parameters		
Total transcripts	**237,314**	**46,690**
N50	**1783**	**1871**
GC content (%)	**0,39**	**0,37**
Transrate v. 1.0.3		
Transrate Assembly Score	**0,0294**	**0,0759**
Transrate Optimal Score	**0,0403**	**0,0812**
Transrate Optimal Cutof	**0,012**	**0,0434**
Good contigs	**217,098**	**42,621**
p good contigs	**0,91**	**0,91**

Table 15 - Evaluation Statistics of rnaSPAdes and Corset Outputs Using Transrate Assembly Validator.

To accurately eliminate redundancies in the assembly, we implemented two filtering steps. Initially, CD-HIT-est was used on the output from rnaSPAdes, and these results were subsequently shared on Figshare (referenced in Table 16).

Label	Name of data file/dataset	File types	Data repository and identifier (DOI or accession number)
Data file 1	**Trinity RNA-Seq de novo transcriptome assembly**	**Fasta file (.fa)**	**https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=3145760812**
Data file 2	**rnaSPADES RNA-Seq de novo transcriptome assembly**	**Fasta file (.fa)**	**https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457605**
Data file 3	Hyla Sarda RNA-Seq de novo transcriptome assembly (Unigenes)	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457614
Data file 4	Open reading frames (ORFs) prediction	Fasta file (.fa)	https://figshare.com/articles/online_resource/Assembly_HS_/17005156?file=31457617
Data file 5	Homology annotation from nonredundant (nr) NCB	Fasta file (.fa)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=3680437234
Data file 6	Homology annotation from SwissProt	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804381
Data file 7	Homology annotation from TrEMBL UniProt	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804405
Data file 8	Homology annotation from nonredundant (nr) protein NCBI	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804348
Data file 9	Homology annotation from SwissProt Protein	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804357
Data file 10	Homology annotation from TrEMBL UniProt Protein	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=36804363
Data file 11	Detailed full Homology annotation Best Hits	Xlsx file (.xlsx)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=368102375
Data file 12	InterProScan results	Text file (.tsv)	https://figshare.com/articles/online_resource/Annotation_HS_/17005198?file=368135224

Table 16 - Summary of data files and datasets generated in this research, including details about the data repository where they are stored.

The final assembly was then created using Corset, which output yielded an N50 of 1871 bp, as shown in Table 15. This dual-step redundancy removal process was crucial in reducing chimeras in the assembly and enhancing the accuracy of further analyses, culminating in the final assembly comprising about 20% of the original transcripts, but with increased N50 value.

The assembly's validation involved two distinct stages. The first stage assessed the preliminary assembly, while the second evaluated the quality of the final, non-redundant assembly. Validation was performed by TransRate and BUSCO, providing a range of metrics

useful in affirming the quality of the de novo assembled transcriptome. The TransRate results (detailed in Table 15) included 'good contigs' figures, indicating the number of high-quality contigs identified by the validator, and the 'p good contigs' value, representing the percentage of high-quality contigs relative to the total contig count in the assembly. We conducted BUSCO evaluations using three orthologous gene databases: Nematoda, Metazoa, and Eukaryota. The completeness of the transcriptome as determined by BUSCO is presented in Table 16.

Busco category	Nematoda	Metazoa	Eukaryota
Complete Buscos (C)	2676 (85,47%	793 (83,12%)	243 (95,30%)
Complete and single-copy Buscos (S)	1703 (54,40%	603 (63,10%)	192 (75,30%)
Complete and duplicated Buscos (D)	970 (31,00%)	190 (20,00%)	51 (20,00%)
Fragmented Buscos (F)	103 (3,30%)	33,39 (3,50%)	8
Missing Buscos (M)	354 (11,30%)	128 (13,40%)	4 (1,60%)
Total Busco groups searched	3131	954	255

Table 16 - Assessment Using BUSCO (v. 5) via gVolante Web Server Applied to Three Databases: Nematoda, Metazoa, and Eukaryota.

Figure 23 illustrates completed, fragmented and missing genes mapped from the three databases.

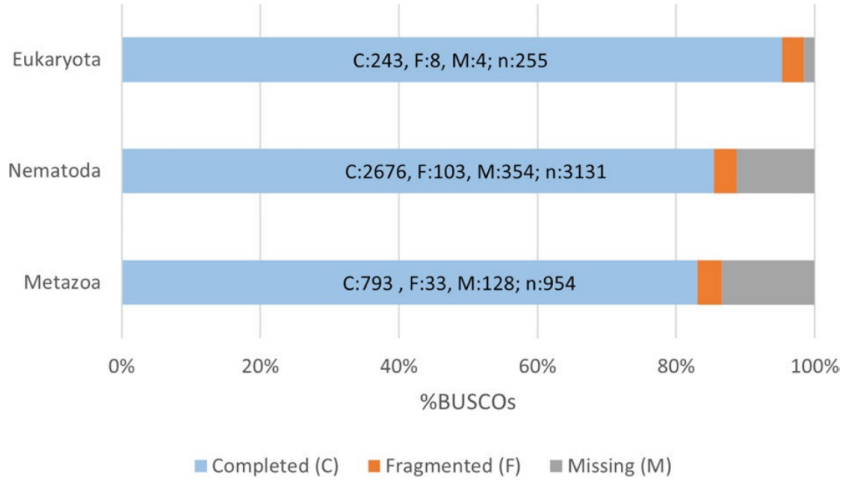
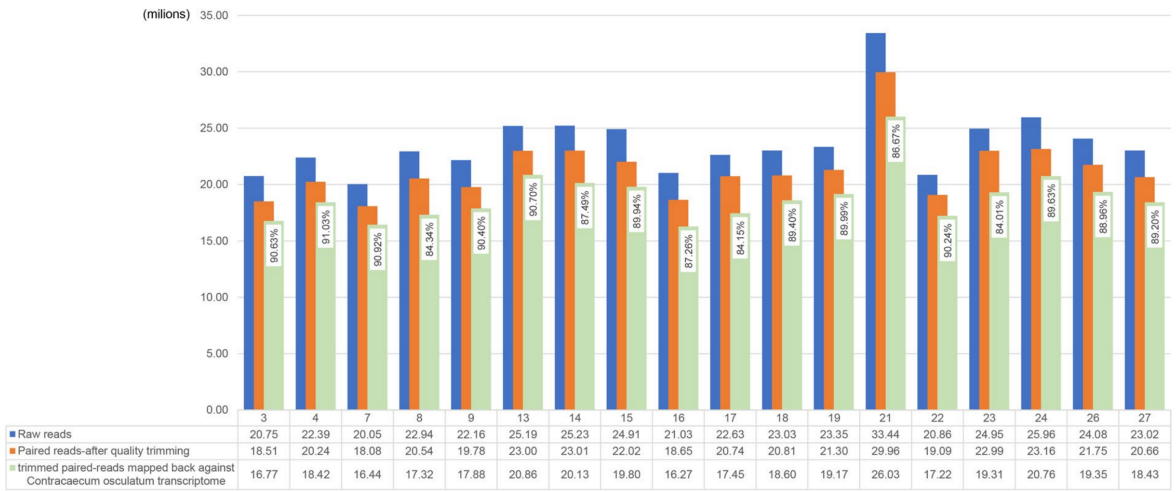


Figure 23 – BUSCO assessment results.

Following the validation and evaluation stage, the assembly data was used as input for the CD-HIT-est program and the output of that, to further refine the final transcriptome dataset, as input for Corset. Output from Corset underwent validation via BUSCO, followed by a quality assessment using HISAT2, mapping the trimmed reads back to the reference transcriptome, composed of unigenes, to evaluate the quality and integrity of the assembled

transcriptome. HISAT2 results showed a percentage of at least 88% (Fig. 24), providing the relative fraction of RNA-Seq reads used to assemble the transcriptome.



	3	4	7	8	9	13	14	15	16	17	18	19	21	22	23	24	26	27
Raw reads	20.75	22.39	20.05	22.94	22.16	25.19	25.23	24.91	21.03	22.63	23.03	23.35	33.44	20.86	24.95	25.96	24.08	23.02
Paired reads-after quality trimming	18.51	20.24	18.08	20.54	19.78	23.00	23.01	22.02	18.65	20.74	20.81	21.30	29.96	19.09	22.99	23.16	21.75	20.66
trimmed paired-reads mapped back against Contracaecum osculatum transcriptome	16.77	18.42	16.44	17.32	17.88	20.86	20.13	19.80	16.27	17.45	18.60	19.17	26.03	17.22	19.31	20.76	19.35	18.43

Figure 24 – For every sample, the data visualization includes three distinct components: the total number of paired-reads is represented in blue, the total paired-reads remaining after adapter removal and quality trimming are depicted in orange, and the trimmed paired-reads that have been successfully mapped against the de novo assembled transcriptome of *Contracaecum osculatum sp. D* are illustrated in green.

The output from Corset was processed using TransDecoder that carries out ORF prediction on both strands of the transcript, steps needed before to proceed with the transcriptome's annotation, performed by DIAMOND against NCBI non-redundant (NR), SwissProt and TrEMBL databases, conducting BLASTX analysis, obtaining annotations for 29,694 (80,3%), 29,904 (80,9%) and 20,660 (55,9%) contigs, respectively. Similarly, using BLASTP versus Nr, TrEMBL and SwissProt, we annotated 24,366 (65,9%), 24,600 (66,5%) and 17,239 (46,6%) contigs, respectively. Results are reported in Table 17.

Database	Number of BLASTX results	Number of BLASTP results
Nr	29,694 (80,30%)	24,366 (65,90%)
TrEMBL	29,904 (80,90%)	24,600 (66,50%)
SwissProt	20,660 (55,90%)	17,239 (46,60%)

Table 17 - Summary of Homology Annotation Hits Across Various Databases Examined in This Study.

The output obtained from the BLASTX annotation consisted of a total of 20,655 sequences mapped simultaneously to the three interrogated databases (i.e., Nr, SwissProt and TrEMBL). The output from the BLASTP annotation consisted of a total of 17,236 sequences mapped simultaneously to the three databases. Venn diagrams are presented in Fig. 4, showing the

redundancy of the annotations in the different databases for both DIAMOND BLASTX (Fig. 25a) and DIAMOND BLASTP (Fig. 25b).

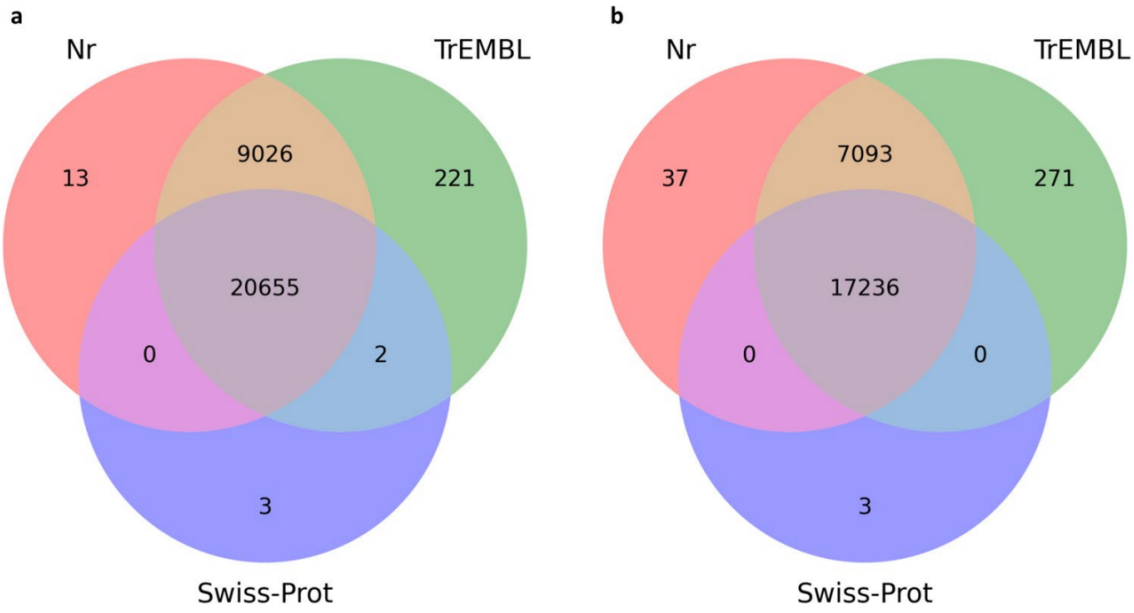


Figure 25 – Venn diagrams for the number of contigs annotated with DIAMOND (BLASTX (a) and BLASTP (b) functions) against the three databases: Nr, SwissProt, TrEMBL.

Furthermore, the ten most represented species and the ten hits of the gene product obtained respectively with BLASTX and BLASTP by mapping the transcripts against the reference database Nr are shown in Figures 26 and 27.

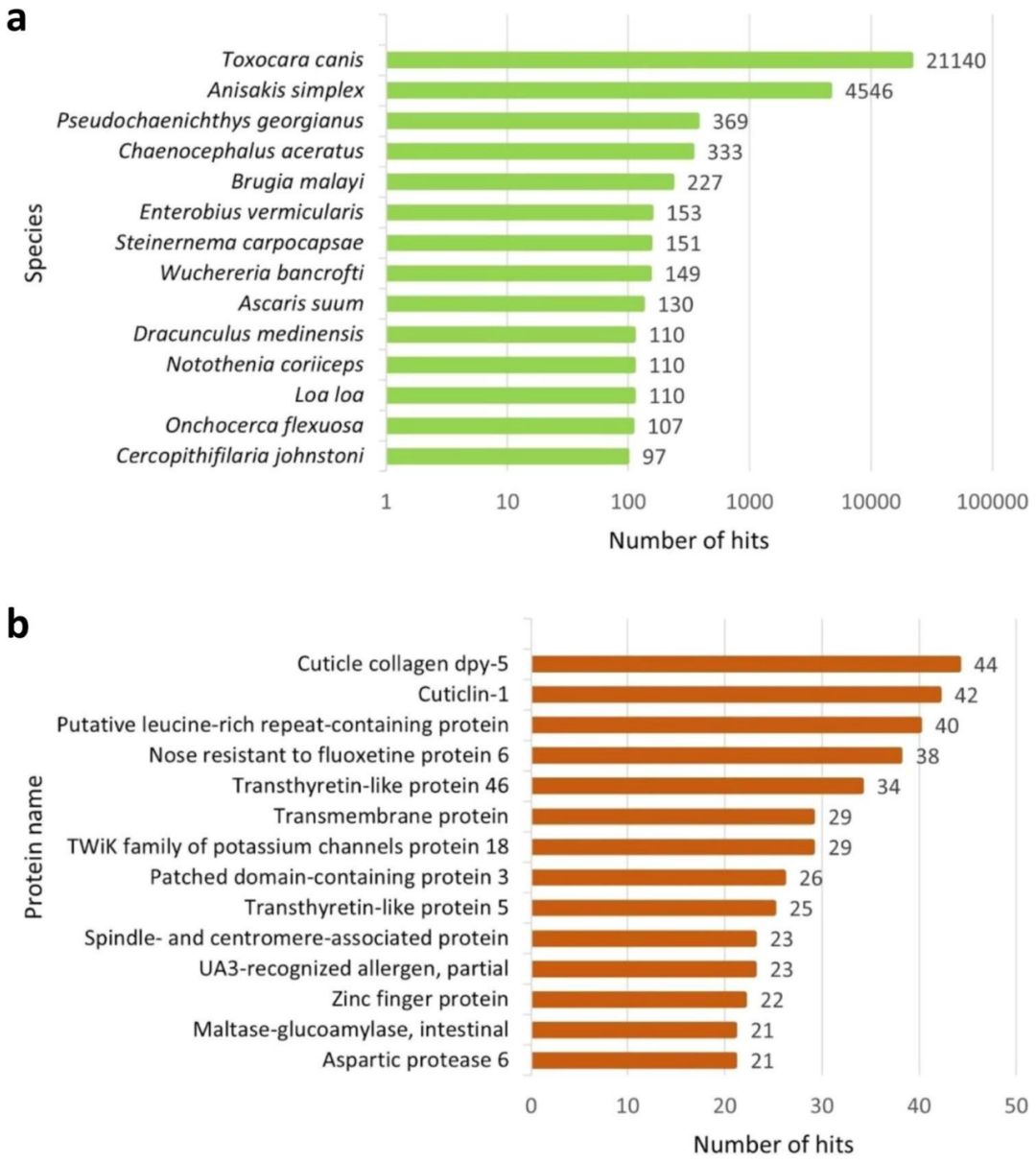


Figure 26 – Most represented species and gene product hits. Top 10 best species (a) and protein (b) hits present in the reference database (Nr, BLASTX).

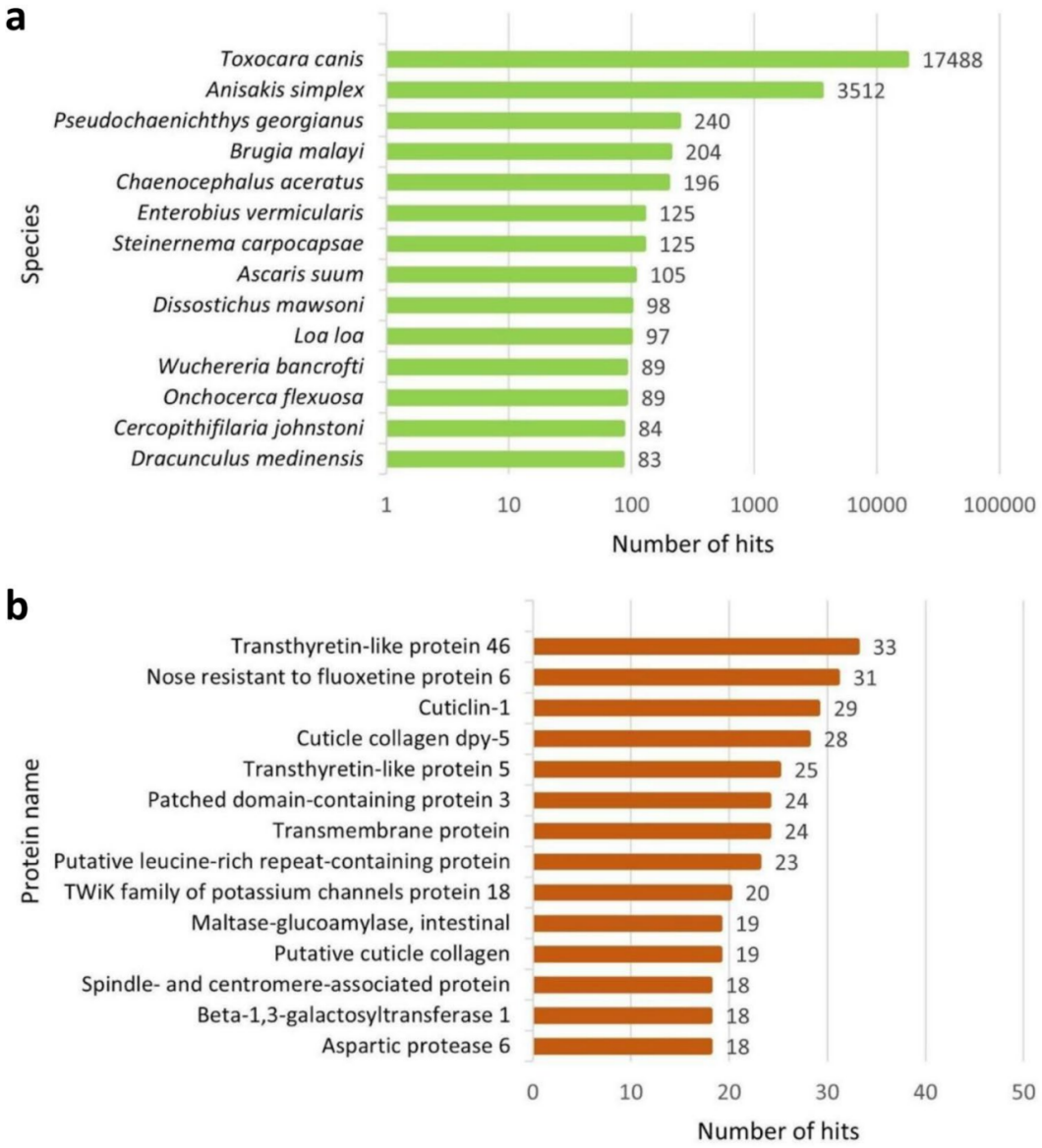


Figure 27 – Most represented species and gene product hits. Top 10 best species (a) and protein (b) hits present in the reference database (Nr, BLASTP).

The total number of unigenes obtained from the transcriptome assembly was also mapped onto another database of functional annotations: EggNOG, that contains detailed functional information for genes within each orthologous group and includes a wide range of sequenced genomes from different species, providing a robust evolutionary context for our data analysis. Of the 36,985 total predicted ORFs, 16,968 (or 45,9%) were annotated in the EggNOG database (results have been stored on Figshare, see Table 16).

Functional annotation of the transcriptome was accomplished by DIAMOND and EggNOG. The application of DIAMOND for annotation purposes led to the identification of 20,655 predicted ORFs (for BLASTX analyses) and 17,236 predicted ORFs (for BLASTP analyses)

shared between the three databases used: Nr, SwissProt and TrEMBL. Finally, from the EggNOG analysis, we obtained COG (Cluster of Orthologous Groups) annotations and KEGG (Kyoto Encyclopedia of Genes and Genomes) annotations for 16,968 ORFs, representing 45,9% of the total.

Comparison with closest species through the orthologs

We compared the predicted ORFs from the de novo transcriptome of C. osculatum sp. D with both the predicted ORFs of the transcriptome of Anisakis pegreffi[99] and the transcripts of A. simplex (sensu stricto). The reference transcriptome of A. simplex (s.s.) was produced with GfRead40. The identification and orthological grouping of all the proteins of the three species were performed using OrthoFinder.

This approach also served to assess the completeness of the assembly based on sequence similarity. OrthoFinder allows orthogroup detection, defined as a set of genes descended from a single gene of the last common ancestor within species groups. The orthogroup detection demonstrated considerable overlap in transcript sequences in all three groups: A. pegreffi, A. simplex (s.s.) and C. osculatum sp. D. More than 20% (8348) of the transcripts identified as putative orthologs were shared between all three species (Figure 28). We found that 15452 transcripts (36,9%) in A. pegreffi, 4049 transcripts (9,7%) in A. simplex (s.s.), and 3378 transcripts (8,1%) in C. osculatum sp. D were classified as species-specific. Tus, the marked level of sequence overlap observed between transcriptomes further validates the completeness and quality of the assembly presented in this study.

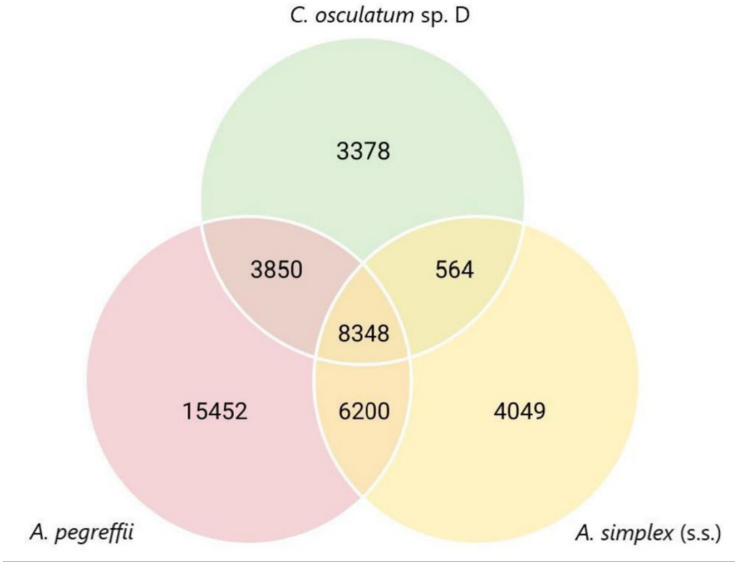


Figure 28 - Venn diagram representing the number of species-specific and overlapping protein orthogroups between the three transcriptomes. The number of *orthogroups* was identified with OrthoFinder.

Chapter 5 - Conclusions

The research on *Bombina pachypus*, *Salamandra salamandra*, *Hyla sarda*, and *Contracaecum osculatum* offers a comprehensive and multifaceted view of biodiversity, adaptation, and ecological dynamics. Each of these non-model organisms, through their unique life histories and ecological niches, has contributed valuable insights into the complex tapestry of life.

Bombina pachypus revealed the intricacies of antipredatory strategies and environmental adaptation, underscoring the genetic underpinnings of behaviors like the 'unken reflex' and aposematic coloration. This research not only illuminated aspects of amphibian biology and ecology but also highlighted the importance of preserving genetic diversity within specific ecological contexts.

Salamandra salamandra provided a window into the genetics of longevity, toxin production, and unique reproductive strategies. The study of this species enriched our understanding of adaptive evolution and ecological interactions, offering lessons on the resilience and adaptability of amphibians in diverse environments. This research also emphasized the need for effective conservation strategies to safeguard these species against environmental threats.

In the case of *Hyla sarda*, our findings shed light on the genetic basis of arboreal adaptation and acoustic communication. This research provided valuable insights into the evolutionary processes shaping these adaptations, especially in the context of island biogeography and the challenges posed by habitat fragmentation and climate change.

Lastly, the study of *Contracaecum osculatum* opened new vistas in understanding the complexities of host-parasite interactions and the genetic mechanisms enabling parasitic survival in diverse marine environments. This work not only deepened our knowledge of marine parasitology but also underscored the ecological significance of parasites in marine food webs and their potential as indicators of ecosystem health.

For each of the species mentioned, a high-quality genomic reference has been meticulously assembled, validated, and published. These transcriptomic datasets serve as valuable resources for further studies, especially in the absence of a fully sequenced and annotated reference genome. They provide critical insights into the genetic underpinnings of various phenotypes and can facilitate a wide range of genomic and evolutionary research. This approach is particularly pivotal for non-model organisms, where genomic resources are often

limited. As such, these de novo transcriptomes lay the foundation for advanced genomic studies and can significantly contribute to our understanding of species biology until comprehensive reference genomes become available.

As we ventured into the realms of assembly and annotation, employing advanced tools like rnaSPAdes, Trinity, and CD-HIT-est, the complexity of transcriptomic data began to unfold. The precision with which these tools have aided in stitching together the fragment of genetic information is nothing short of remarkable. This precision has been instrumental in unraveling the mysteries locked within the RNA sequences, allowing us to glimpse into the functional dynamics of genes in these organisms.

The homology and functional annotations have further enriched our understanding, offering insights into the evolutionary trajectories and adaptive strategies of these species. The use of databases such as NCBI Nr, Swiss-Prot, and TrEMBL, coupled with tools like DIAMOND and InterProScan, has enabled us to map the genetic sequences to known proteins and functions, bridging gaps in our understanding of these non-model species.

The significant findings of this thesis underscore the importance of studying non-model organisms. The diversity they offer presents an untapped reservoir of knowledge that can significantly contribute to our understanding of biology, ecology, and evolution. The methodologies and results discussed herein not only contribute to the specific fields of study for each organism but also enhance the broader scientific community's tools and approaches for genomic research.

In reflecting on this work, it becomes evident that the field of genomics is a mosaic of endless possibilities. Each genetic sequence we decipher adds a piece to the puzzle, enriching our collective understanding of life's complexity. This thesis represents a step forward in this ever-evolving journey, a testament to the power of genomics in unveiling the secrets of nature's diversity.

As we look to the future, the insights gained from this research open new avenues for exploration and discovery. The knowledge acquired here will undoubtedly pave the way for further studies, continuing the legacy of scientific inquiry and advancement. In the grand tapestry of genomic research, this thesis has added vibrant threads, contributing to the ever-growing understanding of the genetic fabric that weaves the tapestry of life on Earth.

In conclusion, the collective findings from these four organisms paint a vivid picture of the richness of biological diversity and the myriad ways in which life adapts and thrives. Our research has not only advanced the frontiers of knowledge in genetics, ecology, and evolutionary biology but also underscored the critical importance of conserving biodiversity. By shedding light on the genetic and ecological dynamics of these species, we contribute to a deeper understanding of the natural world and the pivotal role of non-model organisms in unravelling its mysteries.

Literature Cited

1. Baldridge D, Wangler MF, Bowman AN, Yamamoto S; Undiagnosed Diseases Network; Schedl T, Pak SC, Postlethwait JH, Shin J, Solnica-Krezel L, Bellen HJ, Westerfield M. Model organisms contribute to diagnosis and discovery in the undiagnosed diseases network: current state and a future vision. Orphanet J Rare Dis. 2021 May 7;16(1):206. doi: 10.1186/s13023-021-01839-9. PMID: 33962631; PMCID: PMC8103593.

2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484. PMID: 19015660; PMCID: PMC2949280.

3. Külahoglu C, Bräutigam A. Quantitative transcriptome analysis using RNA-seq. Methods Mol Biol. 2014;1158:71-91. doi: 10.1007/978-1-4939-0700-7_5. PMID: 24792045.

4. Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods. 2011 Apr;8(4 Suppl):S6-11. doi: 10.1038/nmeth.1557. PMID: 21451510; PMCID: PMC3408593.

5. Wolf JB. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour. 2013 Jul;13(4):559-72. doi: 10.1111/1755-0998.12109. Epub 2013 Apr 27. PMID: 23621713.

6. Jensen, P. (2014). Behaviour epigenetics – The connection between environment, stress and welfare. Applied Animal Behaviour Science, 157, 1–7. https://doi.org/10.1016/j.applanim.2014.02.009.

7. van Oers K, Mueller JC. Evolutionary genomics of animal personality. Philos Trans R Soc Lond B Biol Sci. 2010 Dec 27;365(1560):3991-4000. doi: 10.1098/rstb.2010.0178. PMID: 21078651; PMCID: PMC2992743.

8. Van Oers, K. ; Sinn, D.L. / Quantitative and molecular genetics of animal personality. Animal Personalities: Behavior, Physiology, and Evolution. editor / C. Carere ; D. Maestripieri. Chicago; London : The University of Chicago Press, 2013. pp. 149-200

9. Yanofsky CM, Bickel DR. Validation of differential gene expression algorithms: application comparing fold-change estimation to hypothesis testing. BMC Bioinformatics. 2010 Jan 28;11:63. doi: 10.1186/1471-2105-11-63. PMID: 20109217; PMCID: PMC3224549.

10. Gallo CA, Cecchini RL, Carballido JA, Micheletto S, Ponzoni I. Discretization of gene expression data revised. Briefings in Bioinformatics. 2016 Sep;17(5):758-770. DOI: 10.1093/bib/bbv074. PMID: 26438418.

11. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, Sleiman MSB, Ribeiro DM, Bandinelli S, Tanaka T, Nauck M, Völker U, Delaneau O, Metspalu A, Teumer A, Frayling T, Santoni FA, Reymond A, Kutalik Z. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun. 2021 Sep 24;12(1):5647. doi: 10.1038/s41467-021-25805-y. PMID: 34561431; PMCID: PMC8463674.

12. Bisconti R, Canestrelli D, Nascetti G. Genetic diversity and evolutionary history of the Tyrrhenian treefrog Hyla sarda (Anura: Hylidae): adding pieces to the puzzle of Corsica–Sardinia biota. Biological Journal of the Linnean Society. 2011 May 1;103(1):159-67.

13. Bendesky A, Bargmann CI. Genetic contributions to behavioural diversity at the gene-environment interface. Nat Rev Genet. 2011 Nov 8;12(12):809-20. doi: 10.1038/nrg3065. PMID: 22064512.

14. Van Oers, K. & Sinn, D. L. The quantitative and molecular genetics of animal personality. In: Carere, C. & Maestripieri, D. editors. Animal Personalities: Behavior, Physiology, and Evolution. (Chicago: University of Chicago Press; p. 148–200, 2013).

15. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014 Jan;29(1):51-63. doi: 10.1016/j.tree.2013.09.008. Epub 2013 Oct 17. PMID: 24139972.

16. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013 Dec;98(6):236-8. doi: 10.1136/archdischild-2013-304340. Epub 2013 Aug 28. PMID: 23986538; PMCID: PMC3841808.

17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7. doi: 10.1073/pnas.74.12.5463. PMID: 271968; PMCID: PMC431765.

18. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12:323. doi: 10.1186/1471-2105-12-323. PMID: 21816040; PMCID: PMC3163565.

19. Zhao Y, Li MC, Konaté MM, Chen L, Das B, Karlovich C, Williams PM, Evrard YA, Doroshow JH, McShane LM. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med. 2021 Jun 22;19(1):269. doi: 10.1186/s12967-021-02936-w. PMID: 34158060; PMCID: PMC8220791.

20. Zhao Y, Li MC, Konaté MM, Chen L, Das B, Karlovich C, Williams PM, Evrard YA, Doroshow JH, McShane LM. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med. 2021 Jun 22;19(1):269. doi: 10.1186/s12967-021-02936-w. PMID: 34158060; PMCID: PMC8220791.

21. Famoye KF. Count data modeling: Choice between generalized Poisson model and negative binomial model. Journal of Applied Statistical Science. 2005;14(1-2):99-106.

22. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 2019 Dec 23;20(1):295. doi: 10.1186/s13059-019-1861-6. Erratum in: Genome Biol. 2020 Jul 22;21(1):179. PMID: 31870412; PMCID: PMC6927135.

23. Singhal S. De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Mol Ecol Resour. 2013 May;13(3):403-16. doi: 10.1111/1755-0998.12077. Epub 2013 Feb 18. PMID: 23414390.

24. Chiocchio, A., Martino, G., Bisconti, R., Carere, C., Canestrelli D. Shock or jump: deimatic behavior is repeatable and polymorphic in a yellow-bellied toad. bioRxiv 2022.04.29.489992. doi:10.1101/2022.04.29.489992.

25. Umbers KDL, Lehtonen J, Mappes J. Deimatic displays. Curr Biol. 2015 Jan 19;25(2):R58-R59. doi: 10.1016/j.cub.2014.11.011. PMID: 25602301.

26. Arenas LM, Stevens M. Diversity in warning coloration is easily recognized by avian predators. J Evol Biol. 2017 Jul;30(7):1288-1302. doi: 10.1111/jeb.13074. Epub 2017 Apr 21. PMID: 28338250; PMCID: PMC5518184.

27. Richards-Zawacki C, Yeager J, Bart HS. No evidence for differential survival or predation between sympatric color morphs of an aposematic poison frog. Evol Ecol. 2013;27(4):783–95. 10.1007/s10682-013-9636-0.

28. Rönkä, K. Evolution of signal diversity: predator-prey interactions and the maintenance of warning color polymorphism in the wood tiger moth Arctia plantaginis. Jyväskylä studies in biological and environmental science 339 (2017).

29. Joron M, Mallet JL. Diversity in mimicry: paradox or paradigm? Trends Ecol Evol. 1998 Nov 1;13(11):461-6. doi: 10.1016/s0169-5347(98)01483-9. PMID: 21238394.

30. Spadavecchia G, Chiocchio A, Bisconti R, Canestrelli D. Paso doble: A two-step Late Pleistocene range expansion in the Tyrrhenian tree frog Hyla sarda. Gene. 2021 May 15;780:145489. doi: 10.1016/j.gene.2021.145489. Epub 2021 Feb 13. PMID: 33588038.

31. Canestrelli D, Bisconti R, Carere C. Bolder Takes All? The Behavioral Dimension of Biogeography. Trends Ecol Evol. 2016 Jan;31(1):35-43. doi: 10.1016/j.tree.2015.11.004. Epub 2015 Dec 11. PMID: 26688444.

32. Miller TEX, Angert AL, Brown CD, Lee-Yaw JA, Lewis M, Lutscher F, Marculis NG, Melbourne BA, Shaw AK, Szűcs M, Tabares O, Usui T, Weiss-Lehman C, Williams JL. Eco-evolutionary dynamics of range expansion. Ecology. 2020 Oct;101(10):e03139. doi: 10.1002/ecy.3139. Epub 2020 Sep 2. PMID: 32697876.

33. Lindström T, Brown GP, Sisson SA, Phillips BL, Shine R. Rapid shifts in dispersal behavior on an expanding range edge. Proc Natl Acad Sci U S A. 2013 Aug 13;110(33):13452-6. doi: 10.1073/pnas.1303157110. Epub 2013 Jul 29. PMID: 23898175; PMCID: PMC3746873.

34. Pizzatto, L., Both, C., Brown, G., and Shine, R. (2017). The accelerating invasion: dispersal rates of cane toads at an invasion front compared to an already-colonized location. Evol. Ecol. 31, 533–545. doi: 10.1007/s10682-017-9896-1

35. Canestrelli D, Porretta D, Lowe WH, Bisconti R, Carere C, Nascetti G. The Tangled Evolutionary Legacies of Range Expansion and Hybridization. Trends Ecol Evol. 2016 Sep;31(9):677-688. doi: 10.1016/j.tree.2016.06.010. Epub 2016 Jul 20. PMID: 27450753.

36. Werner, E. E. and Gilliam, J. F. (1984) 'The Ontogenetic Niche and Species Interactions in Size-Structured Populations', Annual Review of Ecology and Systematics. Annual Reviews, 15(1), pp. 393–425. doi: 10.1146/annurev.es.15.110184.002141.

37. Ousterhout, B. H. and Semlitsch, R. D. (2018) ‘Effects of conditionally expressed phenotypes and environment on amphibian dispersal in nature’, Oikos. John Wiley & Sons, Ltd, 127(8), pp. 1142–1151. doi: 10.1111/oik.05276.

38. Schulte, U., Küsters, D. and Steinfartz, S. (2007) ‘A PIT tag based analysis of annual movement patterns of adult fire salamanders (Salamandra salamandra) in a Middle European habitat’, Amphibia-Reptilia, 28, pp. 531–536. doi: 10.1163/156853807782152543.

39. Libro, P. et al. (2022) ‘First brain de novo transcriptome of the Tyrrhenian tree frog, Hyla sarda, for the study of dispersal behavior’, Frontiers in Ecology and Evolution, 10. doi: 10.3389/fevo.2022.947186.

40. Chiocchio A, Libro P, Martino G, Bisconti R, Castrignanò T, Canestrelli D. Brain de novo transcriptome assembly of a toad species showing polymorphic anti-predatory behavior. Sci Data. 2022 Oct 13;9(1):619. doi: 10.1038/s41597-022-01724-5. PMID: 36229462; PMCID: PMC9561626.

41. Rey S, Boltana S, Vargas R, Roher N, Mackenzie S. Combining animal personalities with transcriptomics resolves individual variation within a wild-type zebrafish population and identifies underpinning molecular differences in brain function. Mol Ecol. 2013 Dec;22(24):6100-15. doi: 10.1111/mec.12556. Epub 2013 Nov 8. PMID: 24118534.

42. Harris RM, Hofmann HA. Neurogenomics of behavioral plasticity. Adv Exp Med Biol. 2014;781:149-68. doi: 10.1007/978-94-007-7347-9_8. PMID: 24277299.

43. Shin SC, Ahn DH, Kim SJ, Pyo CW, Lee H, Kim MK, Lee J, Lee JE, Detrich HW, Postlethwait JH, Edwards D, Lee SG, Lee JH, Park H. The genome sequence of the Antarctic bullhead notothen reveals evolutionary adaptations to a cold environment. Genome Biol. 2014 Sep 25;15(9):468. doi: 10.1186/s13059-014-0468-1. PMID: 25252967; PMCID: PMC4192396.

44. Kim BM, Amores A, Kang S, Ahn DH, Kim JH, Kim IC, Lee JH, Lee SG, Lee H, Lee J, Kim HW, Desvignes T, Batzel P, Sydes J, Titus T, Wilson CA, Catchen JM, Warren WC, Schartl M, Detrich HW 3rd, Postlethwait JH, Park H. Antarctic blackfin icefish genome reveals adaptations to extreme environments. Nat Ecol Evol. 2019 Mar;3(3):469-478. doi: 10.1038/s41559-019-0812-7. Epub 2019 Feb 25. PMID: 30804520; PMCID: PMC7307600.

45. Lauritano C, Roncalli V, Ambrosino L, Cieslak MC, Ianora A. First De Novo Transcriptome of the Copepod *Rhincalanus gigas* from Antarctic Waters. Biology (Basel). 2020 Nov 23;9(11):410. doi: 10.3390/biology9110410. PMID: 33266516; PMCID: PMC7700397.

46. Timi JT, Poulin R. Why ignoring parasites in fish ecology is a mistake. Int J Parasitol. 2020 Sep;50(10-11):755-761. doi: 10.1016/j.ijpara.2020.04.007. Epub 2020 Jun 24. PMID: 32592807.

47. Mattiucci S, Cipriani P, Paoletti M, Nardi V, Santoro M, Bellisario B, Nascetti G. Temporal stability of parasite distribution and genetic variability values of Contracaecum osculatum sp. D and C. osculatum sp. E (Nematoda: Anisakidae) from fish of the Ross Sea (Antarctica). Int J Parasitol Parasites Wildl. 2015 Oct 22;4(3):356-67. doi: 10.1016/j.ijppaw.2015.10.004. PMID: 26767164; PMCID: PMC4683570.

48. Mattiucci S, Nascetti G. Advances and trends in the molecular systematics of anisakid nematodes, with implications for their evolutionary ecology and host-parasite co-evolutionary processes. Adv Parasitol. 2008;66:47-148. doi: 10.1016/S0065-308X(08)00202-9. PMID: 18486689.

49. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov 8;29(11):987-91. doi: 10.1038/nbt.2023. PMID: 22068540; PMCID: PMC5531759.

50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52. doi: 10.1038/nbt.1883. PMID: 21572440; PMCID: PMC3571712.

51. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012 Dec 27;1(1):18. doi: 10.1186/2047-217X-1-18. Erratum in: Gigascience. 2015;4:30. PMID: 23587118; PMCID: PMC3626529.

52. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr 15;28(8):1086-92. doi: 10.1093/bioinformatics/bts094. Epub 2012 Feb 24. PMID: 22368243; PMCID: PMC3324515.

53. Ponsero AJ, Miller M, Hurwitz BL. Comparison of k-mer-based de novo comparative metagenomic tools and approaches. Microbiome Res Rep. 2023 Jul 20;2(4):27. doi: 10.20517/mrr.2023.26. PMID: 38058765; PMCID: PMC10696585.

54. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021 Sep 27;38(10):4647-4654. doi: 10.1093/molbev/msab199. PMID: 34320186; PMCID: PMC8476166.

55. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform. 2022 Mar 10;23(2):bbab563. doi: 10.1093/bib/bbab563. PMID: 35076693; PMCID: PMC8921630.

56. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res. 2018 Aug 24;7:1338. doi: 10.12688/f1000research.15931.2. PMID: 30254741; PMCID: PMC6124377.

57. de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res. 2019 Nov 7;8:1874. doi: 10.12688/f1000research.21142.2. PMID: 33552473; PMCID: PMC7845152.

58. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience. 2019 Sep 1;8(9):giz100. doi: 10.1093/gigascience/giz100. PMID: 31494669; PMCID: PMC6736328.

59. Wang P, Wang F. A proposed metric set for evaluation of genome assembly quality. Trends Genet. 2023 Mar;39(3):175-186. doi: 10.1016/j.tig.2022.10.005. Epub 2022 Nov 17. PMID: 36402623.

60. Wu B, Li M, Liao X, Luo J, Wu F, Pan Y, Wang J. MEC: Misassembly Error Correction in contigs based on distribution of paired-end reads and statistics of GC-contents. IEEE/ACM Trans Comput Biol Bioinform. 2018 Oct 18. doi: 10.1109/TCBB.2018.2876855. Epub ahead of print. PMID: 30334805.

61. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016 Aug;26(8):1134-44. doi: 10.1101/gr.196469.115. Epub 2016 Jun 1. PMID: 27252236; PMCID: PMC4971766.

62. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014 Dec 21;15(12):553. doi: 10.1186/s13059-014-0553-5. PMID: 25608678; PMCID: PMC4298084.

63. Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min. 2015 Jan 8;8(1):1. doi: 10.1186/s13040-014-0034-0. PMID: 25621011; PMCID: PMC4304608.

64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

65. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417-419. doi: 10.1038/nmeth.4197. Epub 2017 Mar 6. PMID: 28263959; PMCID: PMC5600148.

66. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150-2. doi: 10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142.

67. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov;35(11):1026-1028. doi: 10.1038/nbt.3988. Epub 2017 Oct 16. PMID: 29035372.

68. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. J Vis Exp. 2021 Sep 18;(175). doi: 10.3791/62528. PMID: 34605806.

69. Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF-FINDER: a vector for high-throughput gene identification. Gene. 2002 Jan 9;282(1-2):33-41. doi: 10.1016/s0378-1119(01)00819-8. PMID: 11814675.

70. Nachtweide S, Stanke M. Multi-Genome Annotation with AUGUSTUS. Methods Mol Biol. 2019;1962:139-160. doi: 10.1007/978-1-4939-9173-0_8. PMID: 31020558.

71. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.

72. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D412-D419. doi: 10.1093/nar/gkaa913. PMID: 33125078; PMCID: PMC7779014.

73. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

74. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009 Oct;23(1):205-11. PMID: 20180275.

75. Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 2020 Jan;29(1):28-35. doi: 10.1002/pro.3711. Epub 2019 Aug 29. PMID: 31423653; PMCID: PMC6933857.

76. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. Protein Sci. 2022 Jan;31(1):47-53. doi: 10.1002/pro.4172. Epub 2021 Aug 26. PMID: 34423492; PMCID: PMC8740838.

77. Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; d'Eustachio, P.; Stein, L. (2010). "Reactome: A database of reactions, pathways and biological processes". Nucleic Acids Research. 39 (Database issue): D691–D697. doi:10.1093/nar/gkq1018. PMC 3013646. PMID 21067998.

78. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol. 2007;406:89-112. doi: 10.1007/978-1-59745-535-0_4. PMID: 18287689.

79. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009 Jan;37(Database issue):D211-5. doi: 10.1093/nar/gkn785. Epub 2008 Oct 21. PMID: 18940856; PMCID: PMC2686546.

80. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015 Jan;43(Database issue):D1049-56. doi: 10.1093/nar/gku1179. Epub 2014 Nov 26. PMID: 25428369; PMCID: PMC4383973.

81. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005 Sep 15;21(18):3674-6. doi: 10.1093/bioinformatics/bti610. Epub 2005 Aug 4. PMID: 16081474.

82. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3. Epub 2003 Apr 3. PMID: 12734009.

83. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 2017 May;27(5):768-777. doi: 10.1101/gr.214346.116. Epub 2017 Feb 23. PMID: 28232478; PMCID: PMC5411771.

84. Libro P, Chiocchio A, De Rysky E, Di Martino J, Bisconti R, Castrignanò T, Canestrelli D. De novo transcriptome assembly and annotation for gene discovery in Salamandra salamandra at the larval stage. Sci Data. 2023 May 27;10(1):330. doi: 10.1038/s41597-023-02217-9. PMID: 37244908; PMCID: PMC10224929.

85. Palomba M, Libro P, Di Martino J, Roca-Geronès X, Macali A, Castrignanò T, Canestrelli D, Mattiucci S. De novo transcriptome assembly of an Antarctic nematode for the study of thermal adaptation in marine parasites. Sci Data. 2023 Oct 19;10(1):720. doi: 10.1038/s41597-023-02591-4. PMID: 37857654; PMCID: PMC10587230.

86. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.

87. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

88. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010 Nov;7(11):909-12. doi: 10.1038/nmeth.1517. Epub 2010 Oct 10. PMID: 20935650.

89. Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. 2014 Jul 26;15(7):410. doi: 10.1186/s13059-014-0410-6. PMID: 25063469; PMCID: PMC4165373.

90. Signal, B. and Kahlke, T. (2021) 'Borf: Improved ORF prediction in de-novo assembled transcriptome annotation', bioRxiv. Cold Spring Harbor Laboratory. doi: 10.1101/2021.04.12.439551.

91. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 2015 Jul 13;43(12):e78. doi: 10.1093/nar/gkv227. Epub 2015 Apr 13. PMID: 25870408; PMCID: PMC4499116.

92. Haas, BJ. https://github.com/TransDecoder/TransDecoder

93. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357-60. doi: 10.1038/nmeth.3317. Epub 2015 Mar 9. PMID: 25751142; PMCID: PMC4655817.

94. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016 Sep;11(9):1650-67. doi: 10.1038/nprot.2016.095. Epub 2016 Aug 11. PMID: 27560171; PMCID: PMC5032908.

95. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59-60. doi: 10.1038/nmeth.3176. Epub 2014 Nov 17. PMID: 25402007.

96. Lang F. SWISS-PROT + TREMBL. Trends Genet. 1997 Oct;13(10):417. doi: 10.1016/s0168-9525(97)01258-4. PMID: 9351344.

97. Bisconti R, Canestrelli D, Colangelo P, Nascetti G. Multiple lines of evidence for demographic and range expansion of a temperate species (Hyla sarda) during the last glaciation. Mol Ecol. 2011 Dec;20(24):5313-27. doi: 10.1111/j.1365-294X.2011.05363.x. Epub 2011 Nov 21. PMID: 22097966.

98. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020 Apr 28;9:ISCB Comm J-304. doi: 10.12688/f1000research.23297.2. PMID: 32489650; PMCID: PMC7222033.

99. Palomba M, Libro P, Di Martino J, Rughetti A, Santoro M, Mattiucci S, Castrignanò T. De novo transcriptome assembly and annotation of the third stage larvae of the zoonotic parasite Anisakis pegreffii. BMC Res Notes. 2022 Jun 25;15(1):223. doi: 10.1186/s13104-022-06099-9. PMID: 35752825; PMCID: PMC9233829.