



University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

Kristian Bodulić

**DEVELOPMENT AND APPLICATION OF A
NOVEL TRANSCRIPTOME QUALITY
ASSESSMENT TOOL IN IDENTIFICATION
AND ANNOTATION OF LONG NON-
CODING RNAS IN SPONGES (PORIFERA)**

DOCTORAL THESIS

Zagreb, 2026.



University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

Kristian Bodulić

**DEVELOPMENT AND APPLICATION OF A
NOVEL TRANSCRIPTOME QUALITY
ASSESSMENT TOOL IN IDENTIFICATION
AND ANNOTATION OF LONG NON-
CODING RNAS IN SPONGES (PORIFERA)**

DOCTORAL THESIS

Supervisor:
Prof. Kristian Vlahoviček, PhD

Zagreb, 2026.



University of Zagreb

PRIRODOSLOVNO-MATEMATIČKI FAKULTET
BIOLOŠKI ODSJEK

Kristian Bodulić

**RAZVOJ I PRIMJENA NOVE METODE ZA
PROVJERU KVALITETE
TRANSKRIPTOMA U IDENTIFIKACIJI I
ANOTACIJI DUGIH NEKODIRAJUĆIH RNA
SPUŽVI (PORIFERA)**

DOKTORSKI RAD

Mentor:
Prof. dr. sc. Kristian Vlahoviček

Zagreb, 2026.

This doctoral thesis was conducted within the Bioinformatics Group at the Department of Biology, Faculty of Science, University of Zagreb, under the supervision of Prof. Kristian Vlahoviček, PhD. The thesis was completed as part of the postgraduate doctoral programme in Biology at the Faculty of Science, University of Zagreb.

ACKNOWLEDGMENT

I would like to express my gratitude to my supervisor, Prof. Kristian Vlahoviček, PhD, for his continuous support and guidance throughout the course of this research. His expertise and insightful comments significantly improved both the quality of this thesis and my development as a researcher. I am especially grateful for his patience and availability throughout all stages of this work.

I would also like to thank my colleagues at the University Hospital for Infectious Diseases “Dr. Fran Mihaljević”, which has been my workplace during my doctoral studies. I am especially grateful to Barbara, Josip, Marija, and Ivan for their support, collaboration, and countless discussions. A special thanks goes to Barbara and Josip for our many coffee breaks, SUDs, and great laughs. Barbara, thank you for being the best office roommate I could have hoped for.

I am also grateful to Viktorija, Marija, Filip, Patrik, Karlo, and Aleksandar for encouraging me every step of the way. Thank you for your support and advice, it meant a great deal.

Most importantly, I would like to thank my family. I am especially grateful to my mother and stepfather for their constant love, support, and invaluable advice, which helped me through many challenges along the way. I would also like to thank my brother for his encouragement and continuous support.

I dedicate this dissertation to my mother. Mom, thank you for being the best role model, this achievement is as much yours as it is mine.

SUPERVISOR INFORMATION

Kristian Vlahoviček is a Full Professor of bioinformatics and computational biology at the Faculty of Science, University of Zagreb. He obtained his PhD in bioinformatics and biochemistry in 2000, focusing on the computational prediction of structural properties of DNA. Following his doctoral studies, he spent a decade as a research fellow at the International Centre for Genetic Engineering and Biotechnology in Trieste, Italy. During this period, his work centered around the classification of protein domains using machine learning approaches.

In 2002, he established a computational biology research group at the University of Zagreb, with the goal of integrating bioinformatics into molecular biology education and research. In 2006, supported by an EMBO Young Investigators Programme installation grant, he returned permanently to Croatia and redirected his research towards genomics. His group develops computational tools and applies machine learning methods to address key questions in developmental genomics, cancer genomics, and metagenomics.

He was appointed Full Professor in 2011 at the Faculty of Science, University of Zagreb. Throughout his career, he has held several international academic positions, including adjunct professorships at the University of Oslo and the University of Skövde. From 2008 to 2012, he served as Head of the Division of Biology, overseeing approximately 150 staff members.

Prof. Vlahoviček is deeply engaged in graduate and doctoral education, teaching courses in bioinformatics, algorithms and programming, machine learning and statistics, and genomics. He has supervised more than 30 doctoral and master's students, many of whom have subsequently pursued successful careers at prominent European research institutions and universities.

His scientific output includes over 50 publications in international peer-reviewed journals, along with numerous invited lectures worldwide. His research interests span developmental and differentiation genomics, metagenomics, population genomics, glycomics, and cancer genomics.

In addition to his research and teaching, he actively contributes to the scientific community as a reviewer for numerous journals and international funding bodies, including Horizon Europe, the European Science Foundation, and various national funding agencies. He has also participated in the evaluation of Croatian structural funds programmes.

Prof. Vlahoviček is a Fellow of Academia Europaea and an active member of several professional societies. His service includes roles on the steering committee of the Unity through Knowledge Fund, participation in national and institutional strategic planning initiatives, and membership on the Steering Committee of the Ruđer Bošković Institute.

University of Zagreb

Doctoral thesis

Faculty of Science

Department of Biology

**DEVELOPMENT AND APPLICATION OF A NOVEL TRANSCRIPTOME QUALITY ASSESSMENT TOOL
IN IDENTIFICATION AND ANNOTATION OF LONG NON-CODING RNAs IN SPONGES (PORIFERA)**

KRISTIAN BODULIĆ

Bioinformatics group, Faculty of Science, University of Zagreb

Horvatovac 102a, 10 000 Zagreb

Long non-coding RNAs (lncRNAs) are important genomic regulatory elements, yet their diversity and functions remain poorly understood in early-branching animal lineages. This work combined the development of three computational methods with comparative transcriptomic analyses to characterize lncRNAs across 14 sponge (Porifera) species. The Comprehensive Assessment of Transcript Sequences (CATS) framework was developed to assess transcriptome quality, comprising a reference-free (CATS-rf) and a reference-based (CATS-rb) module. In addition, a genome decontamination pipeline was built to remove foreign sequence from the analysed genomes. Together, these tools enabled the generation of high-confidence lncRNA catalogues across all studied species. Comparative analyses revealed that sponge lncRNAs are generally shorter than protein-coding transcripts, contain fewer exons, exhibit lower isoform diversity, and are expressed at lower levels. Furthermore, a notable enrichment of transposable elements was found within lncRNA loci. Several lncRNAs also exhibited sequence conservation and synteny among closely related species. Additionally, dynamically expressed lncRNAs associated with immune-like responses were detected. Overall, this dissertation establishes a computational framework for transcriptome analysis in non-model organisms while providing a comprehensive resource for lncRNA investigation in Porifera.

174 pages, 83 figures, 18 tables, 162 references, original in: English

Keywords: Transcriptome, Quality control, Long non-coding RNA, Sponges, Contamination

Supervisor: Prof. Kristian Vlahoviček, PhD

Reviewers: Assoc. Prof. Rosa Karlić, PhD

Helena Četković, PhD, scientific adviser

Prof. Damjan Franjević, PhD

Damir Baranašić, PhD, scientific associate

Sveučilište u Zagrebu

Doktorski rad

Prirodoslovno-matematički fakultet

Biološki odsjek

**RAZVOJ I PRIMENA NOVE METODE ZA PROVJERU KVALITETE TRANSKRIPTOMA U
IDENTIFIKACIJI I ANOTACIJI DUGIH NEKODIRAJUĆIH RNA SPUŽVI (PORIFERA)**

KRISTIAN BODULIĆ

Grupa za bioinformatiku, Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu

Horvatovac 102a, 10 000 Zagreb

Duge nekodirajuće RNA (engl. *long non-coding RNA*, lncRNA) predstavljaju važne regulatorne elemente genoma, no njihova su raznolikost i funkcije još uvijek slabo istražene u većini životinjskih koljena. Ovaj rad objedinjuje razvoj triju računalnih alata i transkriptomsku analizu s ciljem karakterizacije lncRNA u 14 vrsta spužvi (Porifera). Razvijeni alat za procjenu kvalitete transkriptoma (CATS, engl. *Comprehensive Assessment of Transcript Sequences*) sastoji se od modula ovisnog o referentnim podacima (CATS-rf) i modula neovisnog o referentnim podacima (CATS-rb). Nadalje, izrađena je metoda za dekontaminaciju kojom su iz analiziranih genoma spužvi uklonjeni genomski ostatci endosimbionata. Zajedno, ovi su alati omogućili izradu pouzdanih kataloga lncRNA analiziranih vrsta spužvi. Komparativne analize pokazale su da su lncRNA spužvi kraće od transkripata koji kodiraju proteine, sadrže manji broj egzona te pokazuju manju raznolikost izoformi i nižu razinu ekspresije. Također, utvrđeno je da su lokusi lncRNA često obogaćeni transpozonomima. Kod znatnog broja lncRNA pronađena je očuvanost primarne strukture i sintenija, osobito među srodnim vrstama spužvi. Nadalje, identificirane su dinamički eksprimirane lncRNA povezane s imunološkim odgovorom spužvi. Zaključno, ova disertacija uspostavlja računalni okvir za analizu transkriptoma nemodelnih organizama i pruža sveobuhvatan temelj za istraživanja lncRNA u koljenu Porifera.

174 stranice, 83 slike, 18 tablica, 162 literaturna navoda, jezik izvornika: engleski

Ključne riječi: Transkriptom, Kontrola kvalitete, Duge nekodirajuće RNA, Spužve, Kontaminacija

Mentor: Prof. dr. sc. Kristian Vlahoviček

Ocjenjivači: Izv. prof. dr. sc. Rosa Karlić

Dr. sc. Helena Četković, znanstvena savjetnica

Prof. dr. sc. Damjan Franjević

Dr. sc. Damir Baranašić, znanstveni suradnik

TABLE OF CONTENTS

1. INTRODUCTION	1
2. LITERATURE REVIEW	3
2.1. RNA Sequencing	3
2.2. Transcriptome Assembly.....	5
2.2.1. De Novo Transcriptome Assembly	5
2.2.2. Genome-guided Transcriptome Assembly	8
2.3. Transcriptome Assembly Quality	8
2.3.1. Common Errors and Quality Evaluation Strategies in Transcriptome Assembly	8
2.3.2. Reference-free Transcriptome Assembly Quality Evaluation	10
2.3.3. Reference-based Transcriptome Assembly Quality Evaluation	11
2.4. Long Non-coding RNAs	11
2.4.1. General Characteristics of Long Non-coding RNAs	12
2.4.2. Regulatory Functions of Long Non-coding RNAs	12
2.4.3. Classification of Long Non-coding RNAs Relative to Protein-coding Genes	13
2.4.4. Evolutionary Origins of Long Non-coding RNAs	14
2.4.5. Evolutionary Conservation Patterns of Long Non-coding RNAs	15
2.4.6. Long Non-coding RNA Identification and Annotation	15
2.5. Sponges (Porifera).....	16
2.5.1. Morphological and Cellular Organization of Sponges	16
2.5.2. Sponge Microbiome: Implications for Genomics.....	17
2.5.3. Sponge Taxonomy	18
2.5.4. Sponge Phylogeny	19
2.5.5. Advances in Sponge Genomics	19
3. MATERIAL AND METHODS	21
3.1. Methodological Overview	21
3.2. CATS: Comprehensive Assessment of Transcript Sequences.....	22
3.2.1. CATS-rf Algorithm Overview	22
3.2.2. CATS-rb Algorithm Overview	29
3.2.3. CATS Benchmarking Library Preparation	33
3.2.4. CATS-rf Benchmarking	38
3.2.5. CATS-rb Benchmarking	40
3.3. Sponge RNA-seq Transcriptome Analysis.....	41
3.3.1. Sponge RNA-seq Library Retrieval	41
3.3.2. Sponge Transcriptome Assembly and Evaluation.....	42
3.4. Genome Decontamination	44
3.4.1. Genome Decontamination Pipeline.....	44
3.4.2. Genome Decontamination Pipeline Validation	50
3.4.3. Sponge Genome Assembly Decontamination	54
3.5. Genome Annotation	54
3.5.1. Repeat Annotation	54
3.6. Long Non-coding RNA Analysis	56
3.6.1. Long Non-coding RNA Identification.....	56
3.6.2. General Characterization of Long Non-coding RNAs	58

3.6.3.	Long Non-coding RNA Splicing Analysis	58
3.6.4.	Long Non-coding RNA Sequence Composition Analysis	59
3.6.5.	Classification of Long Non-coding RNAs Relative to Protein-coding Genes	60
3.6.6.	Analysis of Associations between Long Non-coding RNAs and Transposable Elements	60
3.6.7.	Evolutionary Conservation Analysis of Long Non-coding RNAs	61
3.6.8.	Long Non-coding RNA Expression Analysis	63
3.6.9.	Differential Expression Analysis of Long Non-Coding RNAs in Response to Microbial-Associated Molecular Patterns	63
4.	RESULTS	66
4.1.	CATS Benchmarking	66
4.1.1.	Association of CATS-rf Scores with Assembly Features and Quality	66
4.1.2.	Association of CATS-rf Scores with Assembly Errors	75
4.1.3.	CATS-rf Parameter Robustness Analysis	80
4.1.4.	Association of CATS-rb Scores with Assembly Features and Quality	81
4.2.	Sponge Transcriptome Assembly Evaluation	87
4.2.1.	De Novo Transcriptome Assembly Quality	87
4.2.2.	Genome-guided and Combined Transcriptome Assembly Quality	89
4.3.	Genome Assembly Decontamination	90
4.3.1.	Genome Decontamination Pipeline Validation	90
4.3.2.	Decontamination of the Analysed Sponge Genome Assemblies	94
4.4.	Sponge Genome Annotation	96
4.4.1.	Repeat Annotation	96
4.4.2.	Protein-coding Gene Annotation	98
4.5.	Long Non-coding RNA Analysis	100
4.5.1.	Long Non-coding RNA Identification.....	100
4.5.2.	General Long Non-coding RNA Characteristics.....	100
4.5.3.	Long Non-coding RNA Splicing Analysis	104
4.5.4.	Long Non-coding RNA Sequence Composition	108
4.5.5.	Genomic Organization of Long Non-coding RNAs Relative to Protein-coding Genes	110
4.5.6.	Association between Long Non-coding RNAs and Transposable Elements	113
4.5.7.	Long Non-coding RNA Evolutionary Conservation	117
4.5.8.	Long Non-coding RNA Expression Levels	125
4.5.9.	Case Study: Differential Expression of Long Non-Coding RNAs in Response to Microbial-Associated Molecular Patterns	128
5.	DISCUSSION	137
5.1.	CATS Enables Comprehensive Transcriptome Assembly Evaluation in Non-model Organisms	137
5.2.	Sponge Transcriptome Assembly Required Combined De Novo, Genome-guided, and Clustering Strategies	141
5.3.	Genome Decontamination Is Essential for Reliable Sponge lncRNA Analysis.....	142
5.4.	Repeat and Protein-coding Gene Annotation Revealed Substantial Genome Heterogeneity Among Sponges	145
5.5.	Sponge Long Non-coding RNA Repertoires are Extensive but Strongly Influenced by RNA-seq Sampling	147

5.6.	Sponge Long Non-coding RNAs Share Major Structural Features with lncRNAs from Other Animals.....	148
5.7.	Sponge Long Non-coding RNAs Show Predominantly Canonical Splicing but Reduced Splice-Junction Support	149
5.8.	Sponge Long Non-coding RNAs Differ from Coding Transcripts in Sequence Composition and Exhibit Diverse Genomic Organization	150
5.9.	Transposable Elements Are Associated with Sponge Long Non-coding RNAs	151
5.10.	Sponge Long Non-coding RNAs Show Limited Sequence and Syntenic Conservation	151
5.11.	Sponge Long Non-coding RNA Genes Are Lowly Expressed and Sample-Specific 153	
5.12.	Long Non-coding RNAs as Putative Regulators of Sponge Immune Responses ..	153
6.	CONCLUSION	155
7.	REFERENCES	157
8.	APPENDIX	172
8.1.	CATS Code and Dataset Availability	172
8.2.	Genome Decontamination Pipeline Code and Dataset Availability	172
8.3.	Long Non-coding RNA Analysis	172
9.	CURRICULUM VITAE	174

1. INTRODUCTION

High-throughput RNA sequencing (RNA-seq) has advanced transcriptomics by enabling high-resolution profiling of diverse RNA species. While RNA-seq has traditionally been used to quantify messenger RNA (mRNA) expression, it is increasingly applied to characterize different classes of regulatory non-coding RNAs. The accuracy and interpretability of RNA-seq analyses heavily depend on methodological choices throughout the workflow, from library preparation to downstream bioinformatic processing (Wang et al., 2009; Conesa et al., 2016).

Transcriptome assembly represents a central step in RNA-seq data analysis, involving the reconstruction of full-length transcripts from sequencing reads. This process can be performed either through *de novo* assembly, which relies on sequence overlap information, or through genome-guided assembly, which utilizes read alignment to a reference genome (Conesa et al., 2016; Antoszewski et al., 2025). While these approaches offer complementary strategies for transcript reconstruction, both are affected by challenges such as variable transcript abundance and alternative splicing. These issues are pronounced in *de novo* assembly of short-read data, where they often lead to assembly errors such as transcript fragmentation or redundancy. This can significantly impact downstream analyses, including isoform identification and functional annotation (Grabherr et al., 2011; Li et al., 2014; Smith-Unna et al., 2016). Consequently, robust evaluation of assembly quality is essential to ensure reliable biological interpretation.

Existing transcriptome quality assessment tools rely on either reference-free or reference-based strategies, each offering distinct advantages. Reference-free methods evaluate assemblies by mapping RNA-seq reads back to the assembled transcripts, using read support, coverage patterns, and probabilistic modeling to infer assembly quality. However, their accuracy is often compromised by uneven transcript abundance, sequencing biases, and ambiguous read alignments (Li et al., 2014; Smith-Unna et al., 2016). In contrast, reference-based approaches evaluate structural accuracy by aligning assembled transcripts to a reference genome or transcriptome. While generally more robust in structure validation, these methods depend heavily on the availability of high-quality reference data. As such, they are inherently limited in their ability to assess transcripts from non-model organisms lacking well-annotated genomes or transcriptomes (Bushmanova et al., 2016).

The described limitations of transcriptome evaluation tools are especially important for the study of long non-coding RNAs (lncRNAs), which are typically expressed at low levels and absent from reference annotations. These transcripts have emerged as key regulators of gene expression across multiple levels of cellular organization, acting through diverse mechanisms that include chromatin remodeling, transcriptional regulation, and post-transcriptional control. They can function as molecular scaffolds, decoys, and guides, modulating gene activity in both the nucleus and cytoplasm. Despite their functional importance, lncRNAs exhibit low sequence conservation and high tissue specificity, which complicates their

identification. As a result, their diversity, evolutionary origins, and functional roles remain poorly understood, particularly outside of well-studied model organisms (Statello et al., 2021; Mattick et al., 2023).

Sponges (phylum Porifera) represent one of the earliest-diverging metazoan lineages, providing a unique evolutionary framework for investigating the origins of gene regulatory mechanisms (Wörheide et al., 2012; Botting & Muir, 2018). Comprehensive analyses of several sponge genomes have revealed a surprisingly complex gene repertoire, including many components of key developmental and signaling pathways found in other animal phyla. This includes wide-ranging transcription factor families, cell adhesion molecules, and elements of conserved signaling cascades (Srivastava et al., 2010; Francis et al., 2017; Kenny et al., 2020).

However, genomic analyses of sponges are complicated by their holobiont biology. Sponge tissues contain diverse microbial communities, and genome assemblies may therefore include sequences derived from bacteria, archaea, unicellular eukaryotes, or other associated organisms (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021). Such contamination can interfere with genome annotation and compromise downstream analyses, including transcriptome evaluation, protein-coding gene (PCG) annotation, and lncRNA characterization. Therefore, genome decontamination represents an essential prerequisite for reliable comparative analyses in sponges.

Despite the advances in sponge genomics, lncRNAs in these organisms remain largely unexplored, and their annotation is hindered by both technical and biological challenges. In this context, the first objective of this work is to develop a novel computational tool for transcriptome quality assessment designed to overcome the limitations of existing methods. The proposed framework integrates a reference-free module that uses RNA-seq read support to compute transcript-level quality metrics, along with a reference-based module that uses the reference genome to assess structural accuracy and completeness. The second objective is to develop and apply a genome decontamination pipeline to generate reliable genomic resources for downstream analyses. The third objective is to use these tools to produce high-quality sponge transcriptome assemblies and to perform systematic characterization of lncRNAs across 14 sponge species.

This work is guided by the central hypothesis that lncRNAs in sponges share fundamental similarities in genomic organization with vertebrate lncRNAs, despite deep evolutionary divergence. By integrating methodological development with comparative transcriptomic analysis, this dissertation seeks to advance the computational framework for transcriptome quality evaluation and genome decontamination while significantly expanding the lncRNA repertoire in early-diverging metazoans.

2. LITERATURE REVIEW

2.1. RNA Sequencing

RNA sequencing has emerged as a powerful and widely used technique for studying gene expression, enabling comprehensive analysis of the transcriptome with high resolution (Wang et al., 2009). Although RNA-seq is most frequently applied to mRNA for gene expression quantification, it also enables the analysis of diverse non-coding RNA species, such as lncRNAs, microRNAs (miRNAs), and other regulatory RNA molecules (Arrigoni et al., 2016). Methodological choices in library preparation significantly influence the composition of the sequenced transcriptome and should be carefully considered based on the biological question of interest.

Library preparation is a critical step in RNA-seq workflows, as it converts RNA molecules into a format compatible with high-throughput sequencing while preserving relevant biological information. The process begins with RNA isolation, where total RNA is extracted from cells under conditions that preserve RNA integrity. Because ribosomal RNA (rRNA) constitutes the majority of total RNA, most protocols include a step to either selectively enrich for polyadenylated transcripts (poly(A) selection) or remove rRNA (rRNA depletion). Poly(A) selection is commonly used when focusing on mature mRNA, whereas rRNA depletion is preferred when studying non-coding RNAs, as it provides a more comprehensive representation of the transcriptome (Wang et al., 2009; Conesa et al., 2016). Following RNA enrichment, library preparation workflows diverge depending on the sequencing platform used, most notably between short-read technologies such as Illumina and long-read platforms such as Oxford Nanopore Technologies. These approaches differ in their underlying methodologies and output profiles, necessitating distinct library preparation strategies (Conesa et al., 2016). Given that this work primarily focuses on short-read assemblies, the subsequent sections will describe experimental protocols and bioinformatic workflows for short-read sequencing data.

In Illumina-based RNA-seq workflows (Figure 1), library preparation is optimized for short-read sequencing and requires fragmentation of the input material into fragments typically ranging from 200 to 400 base pairs (bp). A key step in this process is complementary DNA (cDNA) synthesis, in which RNA molecules are converted into DNA using reverse transcriptase and random hexamer primers. This conversion enhances molecular stability and ensures compatibility with downstream sequencing steps (Wang et al., 2009; Conesa et al., 2016; Metzker, 2010). Fragmentation can be performed either before or after cDNA synthesis. Fragmenting RNA prior to reverse transcription generally results in more uniform transcript coverage, whereas fragmentation after cDNA synthesis may introduce biases related to sequence composition and variable reverse transcription efficiency (Wang et al., 2009; Shi et al., 2021; Li & Jiang, 2012). The resulting fragments undergo end repair and adenine tailing to generate compatible ends for adapter ligation. Sequencing adapters are then ligated to both

ends of the cDNA fragments. These adapters contain platform-specific binding sites and may also include index sequences for sample multiplexing. After adapter ligation, libraries are typically amplified by polymerase chain reaction (PCR) to increase the quantity of cDNA available for sequencing. This workflow produces a population of short cDNA fragments that can be sequenced in parallel, enabling high-throughput profiling of RNA expression (Wang et al., 2009; Conesa et al., 2016).

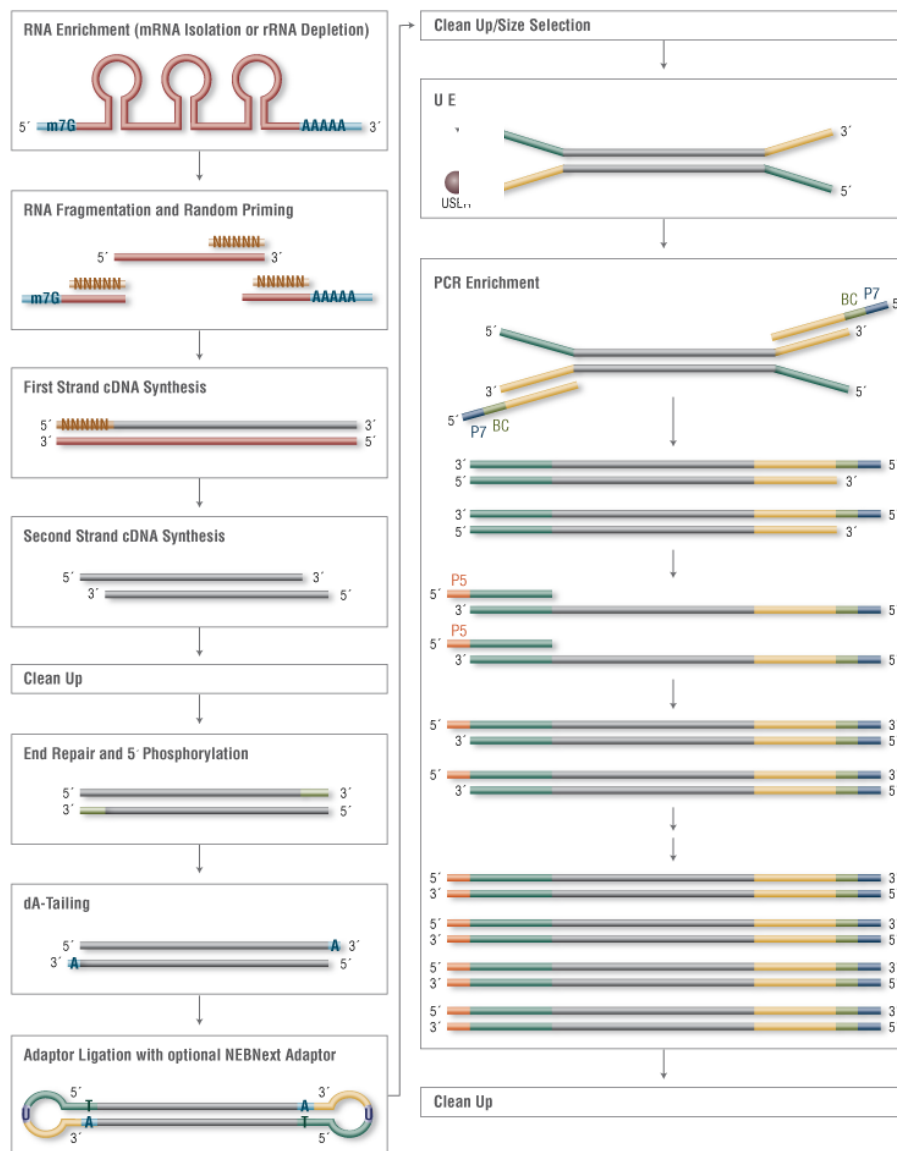


Figure 1. Overview of a typical Illumina RNA-seq library preparation workflow. The protocol includes target RNA enrichment, cDNA synthesis using random priming, end repair and adenylation, adapter ligation, and PCR amplification. mRNA = messenger RNA, rRNA = ribosomal RNA, cDNA = complementary DNA, dA = deoxyadenosine, PCR = polymerase chain reaction. Adapted from New England Biolabs, 2022.

Building on this workflow, Illumina RNA-seq libraries can be sequenced using either single-end or paired-end strategies. In single-end sequencing, fragments are read from one side, providing a single contiguous sequence per fragment, which is generally sufficient for gene expression quantification and is more cost-effective. In contrast, paired-end sequencing reads both ends of each fragment, generating two sequences. This approach facilitates the detection of splice junctions and accurate resolution of transcript isoforms, as the additional positional information between read pairs facilitates transcript reconstruction (Wang et al., 2009; Conesa et al., 2016).

Strand specificity is a notable feature of Illumina RNA-seq library preparation. Stranded RNA-seq protocols preserve information about the original transcriptional orientation, enabling characterization of antisense transcription. This is typically achieved through specific library preparation chemistries that selectively retain strand information. In many protocols, strand specificity is introduced by incorporating modified nucleotides (e.g., deoxyuridine triphosphate, dUTP) during second-strand cDNA synthesis, allowing the second strand to be selectively degraded or not amplified during PCR. As a result, only the first cDNA strand is sequenced, preserving the original transcriptional orientation. In contrast, unstranded protocols do not preserve strand orientation, limiting the analysis of antisense transcription (Parkhomchuk et al., 2009; Levin et al., 2010). Taken together, these considerations highlight that RNA-seq methodologies must be thoughtfully selected in accordance with research objectives.

2.2. *Transcriptome Assembly*

RNA sequencing generates large volumes of reads that require comprehensive bioinformatic processing to extract biologically meaningful information. Initial steps typically include quality control and adapter trimming to ensure data reliability (Conesa et al., 2016; Dawadi et al., 2025). Processed reads can be analysed using two main strategies, depending on the goals of the study. In genome-based approaches, reads are directly aligned to a reference genome, enabling quantification of gene expression levels. This approach is computationally efficient and suited for organisms with well-annotated genomes. Alternatively, transcriptome assembly methods aim to reconstruct transcripts from sequenced reads. Such assembly-based approaches are particularly valuable for identifying novel transcripts and alternative splicing events, thereby providing a more comprehensive view of the analysed transcriptome. Assembly can be performed either *de novo*, where transcripts are reconstructed without a reference genome, or using genome-guided approaches, which leverage an existing reference to improve assembly accuracy (Conesa et al., 2016; Antoszewski et al., 2025).

2.2.1. *De Novo Transcriptome Assembly*

Transcriptome assembly from short-read RNA-seq data presents several computational challenges due to the limited read length. The process is further complicated by the complexity of eukaryotic transcriptomes, which often include variable expression levels,

alternative splicing, transcripts originating from paralogous gene families, and repetitive sequences (Conesa et al., 2016; Martin & Wang, 2011). To address these challenges, specialized de novo assembly algorithms have been developed to reconstruct transcript structures from short reads, commonly employing De Bruijn graph-based approaches (Figure 2). In this framework, sequencing reads are decomposed into shorter subsequences (k-mers), which are connected based on sequence overlap to form a graph structure. The graph is subsequently refined through simplification steps that remove low-confidence elements, enabling the identification of paths corresponding to expressed transcripts. Despite their utility, these methods are computationally demanding and remain susceptible to misassembly (Martin & Wang, 2011).

rnaSPAdes and Trinity are among the most widely used short-read transcriptome assemblers. rnaSPAdes is designed for de novo transcriptome assembly, whereas Trinity can be applied in both de novo and genome-guided contexts (Bushmanova et al., 2019; Grabherr et al., 2011). rnaSPAdes begins by performing read error correction to reduce sequencing noise, followed by constructing multiple De Bruijn graphs at different k-mer sizes to capture transcripts across a wide range of expression levels. Low-abundance transcripts are generally better represented at smaller k values, while highly expressed transcripts are more effectively resolved at larger k values. These graphs are subsequently merged into a unified assembly graph. The integrated graph then undergoes simplification by removing low-coverage edges and collapsing error-induced branches, while retaining paths supported by consistent read coverage. In addition, paired-end read information is actively used to guide path extension and resolve ambiguous connections. Finally, the refined graph is traversed to reconstruct transcript sequences, aiming to recover full-length isoforms while minimizing misassembly (Bushmanova et al., 2019).

De novo assembly by Trinity follows a modular workflow composed of several sequential steps. First, reads are greedily assembled into contiguous sequences by iteratively extending k-mers based on their frequency, selecting the most abundant overlapping k-mers at each step. After assembling contiguous sequences from the most abundant k-mers, those k-mers are removed from the pool and the process is repeated, allowing progressively less expressed transcripts to be reconstructed. Next, related sequences are clustered by analyzing shared k-mer content and incorporating paired-end read information. These clusters group alternatively spliced isoforms or transcripts originating from paralogous genes. For each cluster, a De Bruijn graph is constructed, partitioning the global assembly problem into subgraphs that preserve local transcript complexity. Finally, each graph is processed independently to resolve full-length transcripts. Trinity leverages read support, including coverage consistency and paired-end relationships, to disambiguate branching structures that arise from alternative splicing. Path scoring and pruning strategies are applied to eliminate low-confidence reconstructions. This approach enables Trinity to reconstruct transcripts across a wide range of expression levels (Grabherr et al., 2011).

2.2.2. *Genome-guided Transcriptome Assembly*

Genome-guided transcriptome assemblers, such as Cufflinks and StringTie, take a different approach by reconstructing transcripts from RNA-seq reads that have been aligned to a reference genome. Rather than using De Bruijn graphs, these methods construct splice graphs, where nodes represent exonic regions and edges correspond to splice junctions. Transcript structures are then inferred by traversing the graphs using read coverage and junction support, enabling efficient and structurally accurate reconstruction. However, the performance of these tools is inherently constrained by the quality and completeness of the reference genome and its annotation (Trapnell et al., 2012; Pertea et al., 2015). In contrast, the genome-guided mode of Trinity represents an intermediate approach, where reads are first aligned to the genome to define transcriptional loci, with transcript assembly subsequently performed de novo within these regions. This hybrid strategy improves transcript reconstruction accuracy and isoform resolution while enabling transcript discovery without overly relying on genomic data (Grabherr et al., 2011).

2.3. *Transcriptome Assembly Quality*

The quality of de novo transcriptome assemblies is influenced by multiple factors, including sequencing depth, read length and quality, expression level variability, and the presence of alternative splicing, paralogous sequences, or repeats. These challenges are particularly pronounced in higher eukaryotes, which exhibit greater transcriptomic complexity (Conesa et al., 2016; Martin & Wang, 2011; Li et al., 2016). Among available tools, rnaSPAdes and Trinity have consistently been shown to outperform other de novo assemblers in terms of assembly accuracy and transcript recovery (Hölzer & Marz, 2019; Wang & Gribskov, 2017). However, assembler performance remains highly dependent on parameter selection and dataset-specific characteristics (Hölzer & Marz, 2019; Wang & Gribskov, 2017; Honaas et al., 2016).

2.3.1. *Common Errors and Quality Evaluation Strategies in Transcriptome Assembly*

Transcriptome assembly errors arise from both the technical limitations of short-read sequencing technologies and the intrinsic complexity of transcriptomes (Figure 3). Insertions frequently result from sequencing errors or from incorrect resolution of graph structures, whereby erroneous k-mers are incorporated into contigs during path extension. Another common source of error is transcript collapse, whereby highly similar contigs are merged into a single transcript when shared sequence regions cannot be unambiguously resolved. Structural inaccuracies are also prevalent and include deletions, inversions, and translocations. While deletions typically stem from insufficient coverage across specific regions, inversions and translocations generally result from incorrect traversal of branching paths within assembly graphs or misinterpretation of paired-end read relationships. Fragmentation represents another common assembly error, in which transcripts are reconstructed as multiple partial contigs due to coverage gaps that lead to incomplete graph connectivity. Redundancy is also frequently observed, reflected by multiple highly similar or

partially overlapping contigs derived from a single transcript. Finally, chimeric transcripts are formed through the erroneous joining of sequences originating from distinct genes. Such errors typically arise from shared k-mers that lead to incorrect merging of unrelated sequences (Grabherr et al., 2011; Li et al., 2014; Smith-Unna et al., 2016). Collectively, the described errors underscore the challenges of reconstructing transcriptomes from short-read data and emphasize the necessity for comprehensive evaluation frameworks.

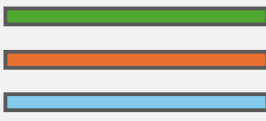
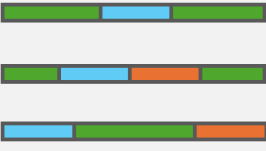
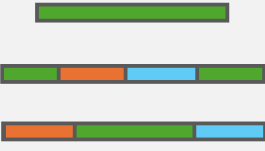
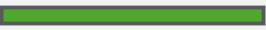
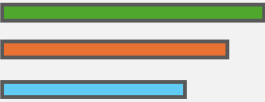
	Native	Error
Insertion		
Transcript collapsing		
Structural errors		
Fragmentation		
Redundancy		
Chimerism		

Figure 3. Common assembly errors in transcriptomes derived from short-read RNA-seq data.

Considering that transcriptome assembly quality can vary substantially depending on both biological and technical factors, robust quality assessment is essential. Several tools have been developed to evaluate assemblies using distinct approaches. Reference-free methods assess assembly quality based on RNA-seq read support, coverage profiles, and probabilistic models without requiring a reference genome (Li et al., 2014; Smith-Unna et al., 2016). In contrast, reference-based tools compare assembled transcripts to a reference genome or transcriptome to assess structural accuracy and completeness (Li et al., 2014; Smith-Unna et al., 2016; Bushmanova et al., 2016). A third complementary approach is exemplified by BUSCO, which evaluates assembly completeness by quantifying the presence of conserved single-copy orthologs (Seppey et al., 2019).

2.3.2. *Reference-free Transcriptome Assembly Quality Evaluation*

Two widely used reference-free tools for transcriptome assembly evaluation are RSEM-EVAL and TransRate, both of which evaluate quality by aligning the RNA-seq reads used during assembly back to the reconstructed transcripts (Li et al., 2014; Smith-Unna et al., 2016). RSEM-EVAL computes a composite assembly score that integrates three components. It first estimates the likelihood of the observed RNA-seq reads given the assembled transcripts using an expectation–maximization algorithm that probabilistically assigns reads to transcripts. RNA-seq reads are initially aligned to the analysed transcripts to generate candidate alignments. During the expectation step, posterior probabilities are calculated for each read–transcript pair, proportional to transcript abundance and alignment quality. In the maximization step, transcript abundance parameters are updated to maximize the expected data likelihood. These steps are iterated until convergence, reaching a local optimum of the likelihood function. Second, RSEM-EVAL incorporates a prior on the assembly that penalizes overly complex transcripts. Third, the tool applies a Bayesian Information Criterion penalty to discourage overfitting by balancing model complexity against goodness of fit. Overall, the RSEM-EVAL assembly score integrates three key aspects: data likelihood, structural plausibility, and parsimony. In addition to the global assembly score, RSEM-EVAL also reports transcript-level scores, which quantify each transcript’s contribution to the overall likelihood and reflect relative transcript quality within a given RNA-seq dataset (Li et al., 2014).

The second reference-free method for evaluating de novo transcriptome assemblies is TransRate. It relies on empirical evidence obtained by aligning RNA-seq reads back to the assembled transcripts, with multimapped reads probabilistically assigned in proportion to transcript abundance. TransRate computes several transcript-level metrics to assess assembly quality. The coverage score quantifies the fraction of bases that remain uncovered by reads, enabling the detection of insertions and redundant transcripts. The nucleotide score reflects the mean edit distance between reads and transcripts, capturing discrepancies that may arise from errors such as collapsed isoforms or repeats. The order score aggregates the consistency of paired-end read alignments, identifying structural errors. The segmentation score examines per-base coverage profiles to detect irregular patterns that may indicate chimeric transcripts. TransRate integrates these metrics into a transcript score by taking their product. The final assembly score is calculated as the geometric mean of all transcript scores, multiplied by the proportion of reads that map back to the assembly. Through this approach, TransRate provides an interpretable assessment of assembly accuracy and completeness given the input reads (Smith-Unna et al., 2016).

Despite their utility, both RSEM-EVAL and TransRate exhibit several notable limitations. In RSEM-EVAL, transcript-level scores are inherently difficult to interpret with respect to specific error types, as they represent contributions to a global likelihood rather than direct measures of transcript accuracy. Furthermore, its probabilistic framework may introduce bias toward overly simplified assemblies due to penalties imposed on model complexity. The algorithm is

also permissive of assemblies that erroneously merge low-coverage transcripts, leading to inflated scores for structurally implausible reconstructions. An additional limitation lies in the employed alignment model, which does not account for insertions and deletions, thereby reducing its sensitivity to several classes of assembly errors (Li et al., 2014). TransRate partially addresses these limitations by leveraging alignment-based evidence to detect specific error types. However, it tends to underestimate the quality of lowly expressed isoforms due to the probabilistic assignment of multimapped reads. Additionally, the independent assignment of multimapped paired-end reads may artificially lower the order score. Furthermore, TransRate exhibits limited ability to distinguish base mismatches caused by low read quality or transcript variation from those caused by misassembly (Smith-Unna et al., 2016). Finally, both tools appear to be no longer maintained, which may limit their applicability in current workflows (Li, 2014; Smith-Unna, 2016).

2.3.3. Reference-based Transcriptome Assembly Quality Evaluation

Reference-based evaluation of de novo transcriptome assemblies involves comparing reconstructed transcripts to a reference transcriptome or genome, thereby providing a more direct assessment of quality compared to reference-free tools. Two notable evaluation methods that utilize reference transcriptomes are REF-EVAL, which is packaged together with RSEM-EVAL, and the reference-based evaluation method implemented in TransRate. Both methods leverage alignment information to compare assembled and reference transcripts, enabling the identification of structural errors and incomplete reconstructions. Nevertheless, the effectiveness of this approach is constrained by the completeness and quality of the reference transcriptome, biasing evaluation against poorly annotated transcripts (Li et al., 2014; Smith-Unna et al., 2016).

RNAQuast extends this framework by evaluating the assembled transcripts utilizing a reference genome. The tool computes a range of metrics related to transcript mappability and assembly completeness, such as the number of transcripts that align to the genome, the fraction of transcripts covering known genes or exons, and measures of structural errors, chimerism, and fragmentation. While these metrics provide valuable insights into assembly quality, RNAQuast does not combine them into a unified score. Furthermore, a key limitation of RNAQuast is its strong dependence on existing genomic annotations (Bushmanova et al., 2016). More broadly, the heavy reliance of reference-based approaches on available reference data reduces their effectiveness for analyzing non-coding transcripts, which are often poorly annotated or entirely absent from reference databases, particularly in non-model organisms.

2.4. Long Non-coding RNAs

While only a limited fraction of the genome encodes proteins, a substantial proportion is nonetheless transcribed, giving rise to a diverse array of non-coding transcripts. Among these, lncRNAs constitute a heterogeneous class of RNA molecules that play important roles in

regulating gene expression. Once dismissed as transcriptional noise, lncRNAs are now recognized as key regulators of biological processes, contributing to chromatin organization, transcriptional modulation, and post-transcriptional control. Their widespread expression underscores the significance of the non-coding parts of the genome (Statello et al., 2021; Mattick et al., 2023).

2.4.1. General Characteristics of Long Non-coding RNAs

Long non-coding RNAs are conventionally defined as RNA transcripts longer than 200 bp that lack significant protein-coding potential and are distinct from well-characterized structural non-coding RNAs. Most lncRNA genes exhibit properties similar to PCGs, including the presence of CpG islands, transcription by RNA polymerase (RNAP) II, exon-intron structures, and alternative splicing. Additionally, many lncRNA transcripts share features with mRNAs, including 5' capping with 7-methylguanosine and 3' polyadenylation. However, a subset of lncRNAs deviates from this pattern, as they are transcribed by RNAP I or RNAP III and do not undergo canonical post-transcriptional modifications. Compared to mRNAs, lncRNAs are typically shorter in length and on average contain fewer exons. Most lncRNA genes are relatively lowly expressed, and it is estimated that the abundance of lncRNA transcripts is generally several times lower than that of mRNA transcripts (Statello et al., 2021; Mattick et al., 2023). Furthermore, the majority of human lncRNAs are tissue-specific, with the greatest diversity observed in the reproductive system and brain (Briggs et al., 2015; Derrien et al., 2012). The expression of individual lncRNA transcripts is highly dependent on the developmental stage, as well as on conditions induced by environmental factors, such as activation of the immune response and hypoxia (Statello et al., 2021; Mattick et al., 2023; Atianand & Fitzgerald, 2014; Peng et al., 2021).

Mutations and dysregulated expression of lncRNA genes have been linked to several diseases, including various cancers and Alzheimer's disease (Statello et al., 2021; Mattick et al., 2023; Bhan et al., 2017; Wei et al., 2018). In addition to their role in disease, the importance of lncRNAs is further underscored by their diversity. Notably, a large proportion of lncRNAs remain functionally uncharacterized, reflecting their complexity and emerging nature. In humans, 35899 lncRNA-encoding genes have been described by 2025, significantly exceeding the number of annotated PCGs (GENCODE, 2025). A similar pattern is observed in a wide range of taxa, including other vertebrates (Darbellay & Necselea, 2020), invertebrates (Jing et al., 2025; Nam & Bartel, 2012), fungi (Avalos et al., 2025), and plants (Jha et al., 2020), highlighting the importance of efficient annotation of lncRNAs across diverse taxonomic groups.

2.4.2. Regulatory Functions of Long Non-coding RNAs

Long non-coding RNAs exert their regulatory functions through diverse mechanisms that operate at multiple levels of gene expression (Figure 4). In the nucleus, lncRNAs can modulate chromatin structure by recruiting chromatin-remodeling complexes, such as histone modifiers and DNA methyltransferases. Through these interactions, lncRNAs can influence chromatin

organization and the accessibility of regulatory regions. At the transcriptional level, lncRNAs can control gene expression by interacting with transcription factors, modulating enhancer–promoter communication, or affecting RNAP recruitment, pausing, and elongation. They can also act as key regulators of post-transcriptional processes. Some lncRNAs function as molecular “sponges” by binding and sequestering miRNAs, thereby relieving repression of miRNA targets. In addition, certain cytoplasmic lncRNAs function as molecular scaffolds, bringing together post-transcriptional regulators into ribonucleoprotein complexes. Other lncRNAs regulate mRNA maturation by influencing alternative splicing through interactions with splicing factors. Specific lncRNAs may also regulate mRNA stability via associations with RNA-binding proteins. Furthermore, lncRNAs can control translation by interacting with ribosomes or translation factors, leading to either enhancement or inhibition of protein synthesis. Finally, lncRNAs can promote or inhibit mRNA degradation by directing transcripts to distinct decay pathways (Statello et al., 2021; Mattick et al., 2023; Salehi et al., 2017).

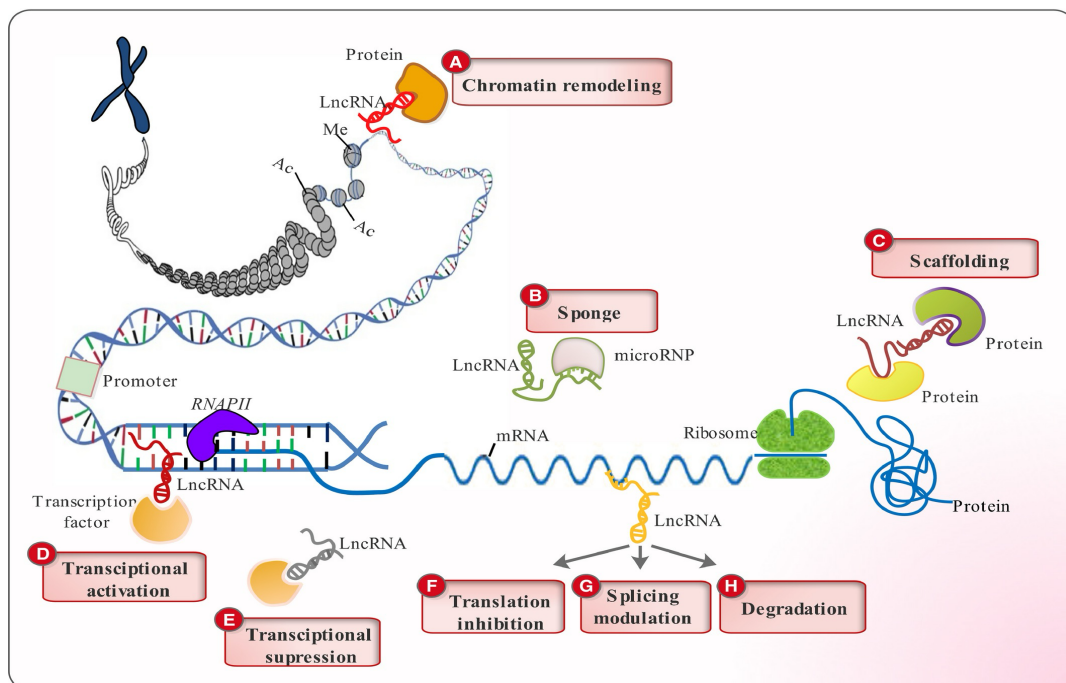


Figure 4. Mechanisms of lncRNA-mediated gene regulation, including chromatin remodeling, transcriptional modulation, and post-transcriptional control. mRNA = messenger RNA, lncRNA = long non-coding RNA, RNAP = RNA polymerase. Reproduced from Salehi et al., 2017.

2.4.3. Classification of Long Non-coding RNAs Relative to Protein-coding Genes

Long non-coding RNAs are generally classified according to their genomic position and orientation relative to PCGs (Figure 5). Intergenic lncRNAs are transcribed from regions located between PCGs, whereas intronic lncRNAs originate entirely from introns of PCGs. In contrast, sense and antisense lncRNAs overlap exonic regions of PCGs, with sense lncRNAs transcribed from the same strand and antisense lncRNAs originating from the opposite strand. Bidirectional lncRNAs are transcribed in the opposite direction from nearby PCGs with which

they typically share a promoter region. Additionally, overlapping lncRNAs represent a more complex category, encompassing transcripts that may combine features of the described lncRNA classes. This classification is linked to lncRNA function, as their genomic context often influences their regulatory roles, particularly in the modulation of neighboring PCG expression in cis (Liu et al., 2021).

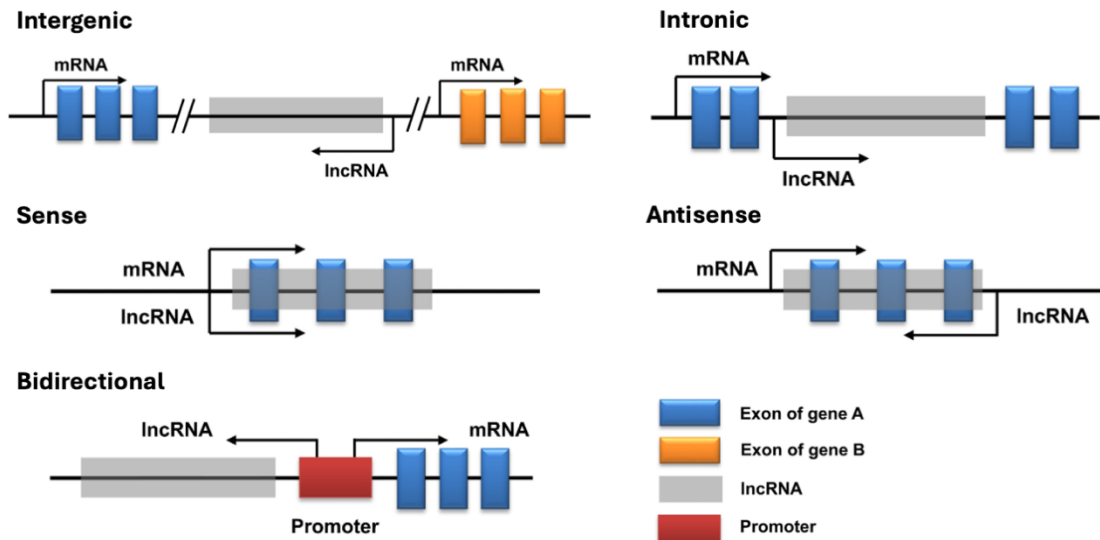


Figure 5. Genomic organization of lncRNAs relative to PCGs. The employed classification includes intergenic, intronic, sense, antisense, and bidirectional lncRNAs. mRNA = messenger RNA, lncRNA = long non-coding RNA. Adapted from Liu et al., 2021.

2.4.4. Evolutionary Origins of Long Non-coding RNAs

The origin of lncRNAs is closely associated with multiple evolutionary mechanisms. One prominent pathway involves the duplication of PCGs, followed by the gradual accumulation of mutations in one gene copy. Over time, this process can give rise to pseudogenes, genomic sequences that have lost their protein-coding capacity but often retain substantial sequence similarity to their parental genes. Although pseudogenes are not translated into proteins, many are transcriptionally active and give rise to lncRNA transcripts. In some cases, extensive mutational divergence results in pseudogenes that no longer resemble their ancestral coding genes (Mattick et al., 2023; Milligan & Lipovich, 2015). Similarly, lncRNAs can also arise through duplication events of pre-existing lncRNA genes, followed by sequence diversification and functional divergence (Hezroni et al., 2015).

Beyond pseudogene formation, many lncRNA genes arise through genomic rearrangements, reflecting the inherently modular organization of lncRNA molecules. These transcripts are typically composed of distinct functional domains, each contributing to specific interactions. For example, specific domains can adopt secondary structures that mediate protein interactions, whereas others function through primary sequence complementarity to target DNA or RNA molecules. Furthermore, lncRNA genes often contain regulatory elements such

as promoters, enhancers, and silencers, which control their expression patterns (Statello et al., 2021; Mattick et al., 2023). The assembly of these structural and regulatory modules through genomic rearrangements is therefore a key driver in the emergence of new lncRNA genes.

A major contributor to such rearrangements is the activity of transposable elements (TEs) (Mattick et al., 2023; Hezroni et al., 2015; Kapusta et al., 2013). These mobile genetic elements are broadly classified into two groups: retrotransposons, which propagate via an RNA intermediate through a copy-and-paste mechanism, and DNA transposons, which typically move through a cut-and-paste mechanism. Retrotransposons include subclasses such as long terminal repeat (LTR) retrotransposons, long interspersed nuclear elements (LINEs), and short interspersed nuclear elements (SINEs) (Kazazian, 2004). A distinct class of DNA transposons, known as helitrons, transposes via a rolling-circle replication mechanism and is notable for its ability to capture and mobilise fragments of host genes (Thomas & Pritham, 2015). Notably, lncRNA genes are often highly enriched in TE sequences compared to PCGs. This enrichment supports the hypothesis that insertion of TEs into non-coding regions can lead to de novo formation of lncRNA genes or introduce new functional domains into existing genes (Mattick et al., 2023; Hezroni et al., 2015; Kapusta et al., 2013). In addition, TEs embedded within lncRNA exons may contribute to the stabilization of their secondary structure, thereby enhancing their functional capacity (Kapusta et al., 2013). Altogether, these observations highlight TEs as central players in both the origin and functional diversification of lncRNAs.

2.4.5. Evolutionary Conservation Patterns of Long Non-coding RNAs

A further important aspect of lncRNA biology concerns their sequence homology and synteny across species. In contrast to PCGs, lncRNAs generally exhibit low primary sequence conservation, showing limited similarity even among closely related species. However, this lack of sequence conservation does not necessarily imply functional insignificance. In many cases, lncRNAs display conservation at the structural or functional level, with preserved secondary structures or interaction domains enabling similar molecular roles across taxa. Moreover, synteny has emerged as a key indicator of lncRNA conservation. Numerous lncRNAs retain conserved genomic positions despite sequence divergence, suggesting that their regulatory functions, particularly in cis, may depend more on genomic context than on precise sequence identity (Hezroni et al., 2015; Huang et al., 2024; Quinn et al., 2016). These findings highlight the importance of integrating multiple levels of conservation when assessing lncRNA function and evolution.

2.4.6. Long Non-coding RNA Identification and Annotation

The characterization of lncRNAs presents substantial challenges across taxa, largely due to their low sequence conservation, variable expression levels, and frequent overlap with other genomic features (Mattick et al., 2023; Ramakrishnaiah et al., 2020). In model organisms such as *Homo sapiens* and *Mus musculus*, extensive transcriptomic resources and high-quality

reference genomes have enabled the development of comprehensive lncRNA catalogues (GENCODE, 2025). These annotations rely on pipelines that combine high-throughput RNA-seq, transcriptome assembly, and coding potential prediction algorithms. Despite these advances, lncRNA annotation remains less complete and more error-prone than that of PCGs, even in model systems. Challenges include distinguishing bona fide lncRNAs from transcriptional noise, accurately reconstructing lowly expressed transcripts, and resolving overlapping transcription (Mattick et al., 2023; Ramakrishnaiah et al., 2020). In non-model organisms, these challenges are amplified due to limited genomic resources and the absence of curated gene annotations (Ungaro et al., 2017; Singhal, 2013). As a result, lncRNA identification in such taxa often relies heavily on de novo transcriptome assembly, with reduced confidence in transcript boundaries and functional predictions. Nonetheless, expanding research efforts in diverse lineages have begun to uncover the evolutionary breadth of lncRNAs beyond traditional model organisms.

Early-branching metazoans, particularly sponges, have emerged as valuable systems for studying the origin and evolution of lncRNAs. Investigations in several sponge species have revealed unexpectedly rich repertoires of lncRNA transcripts. In the model sponge *Amphimedon queenslandica*, genome-wide analyses have identified numerous lncRNAs exhibiting developmental stage-specific expression, suggesting roles in regulating early metazoan developmental processes (Gaiti et al., 2015). Similarly, a study in *Sycon ciliatum* has demonstrated dynamic lncRNA expression patterns associated with cell differentiation and tissue organization (Fortunato et al., 2014). More recently, a comprehensive analysis in *Eunapius subterraneus* identified a diverse set of lncRNAs, further expanding the known repertoire in sponges (Bodulić, 2020). Notably, lncRNAs from all three species share several features with those of bilaterian animals, including genomic organization, TE enrichment, and regulated expression, yet display minimal sequence conservation (Gaiti et al., 2015; Fortunato et al., 2014; Bodulić, 2020). Despite these advances, lncRNAs remain largely unexplored in the vast majority of sponge species, highlighting a significant gap in the understanding of their diversity and functional roles.

2.5. Sponges (Porifera)

Sponges (phylum Porifera) are among the earliest-diverging lineages of animals and provide key insights into the evolution of metazoans (Wörheide et al., 2012; Botting & Muir, 2018). They are aquatic organisms characterized by a porous body structure and filter-feeding lifestyle. Despite lacking true tissues and organs, sponges possess a diverse set of cell types and a relatively complex genetic toolkit (Srivastava et al., 2010; Leininger et al., 2014).

2.5.1. Morphological and Cellular Organization of Sponges

Sponge morphology is characterized by a relatively simple body plan organized around an aquiferous system of canals and chambers that facilitates efficient water circulation. Water enters the body through microscopic pores (ostia) formed by porocytes and exits via larger

openings, the oscula, whose diameter can be regulated by contractile myocytes. The internal cavity (spongocoel) is lined with choanoderm, where the flagellar activity of choanocytes generates water currents and enables the capture and phagocytosis of suspended food particles. The body wall consists of an outer pinacoderm composed of flattened pinacocytes, which provide protection and limited contractile function. The pinacoderm and choanoderm are separated by the mesohyl, a gelatinous extracellular matrix containing a diverse array of specialized cells. Amoebocytes play key roles in intracellular digestion, nutrient transport, and cellular differentiation, whereas sclerocytes are responsible for the secretion of mineral spicules. Spongocytes produce the fibrous protein spongin, contributing to the skeletal framework that, together with mineralized elements, provides structural support. Notably, both the organization of the aquiferous system and the composition of skeletal elements vary among sponge classes (Simpson, 1984; Leys & Hill, 2012).

2.5.2. Sponge Microbiome: Implications for Genomics

Sponges host dense and taxonomically diverse microbial communities that are integral to their biology (Figure 6). These microbiomes comprise bacteria, archaea, viruses, and microbial eukaryotes, and can account for up to 40% of the sponge's total biomass. Microbial assemblages are often host-specific and temporally stable, indicating long-term coevolutionary relationships. Functionally, sponge-associated microbes contribute to a wide range of metabolic processes, including carbon and nitrogen fixation, sulfur metabolism, and phosphorus turnover. Many symbionts produce secondary metabolites, including antimicrobial and cytotoxic compounds that protect the host from predators and pathogens. This holobiont framework emphasizes that sponge physiology and ecological performance emerge from the collective interactions between host and microbiome (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021).

At the same time, the association with diverse microorganisms complicates genomic and transcriptomic analyses, as sequenced data often contain a mixture of host and symbiont-derived material. This necessitates careful filtering and annotation strategies to accurately disentangle sponge-derived sequences from those of associated microbes, which can otherwise lead to erroneous interpretations (Steffen et al., 2023; Kenny et al., 2020). Several approaches have been developed to address contamination in genome assemblies, including composition-based methods, coverage-based clustering, marker gene detection, and taxonomic classification involving protein- or nucleotide-based similarity. Although effective in many contexts, these methods are often applied as single-strategy frameworks and do not integrate multiple complementary signals into a unified scoring scheme (Astashyn et al., 2024; Bálint et al., 2024; Lataretu et al., 2025). Furthermore, these tools have not been systematically evaluated on non-model metazoans such as sponges, which are characterized by poorly defined symbiotic communities, limited representation in reference databases, and extensive horizontal gene transfer (HGT), potentially leading to erroneous contamination calls (Hentschel et al., 2012; Pita et al., 2018; Kelly et al., 2022).

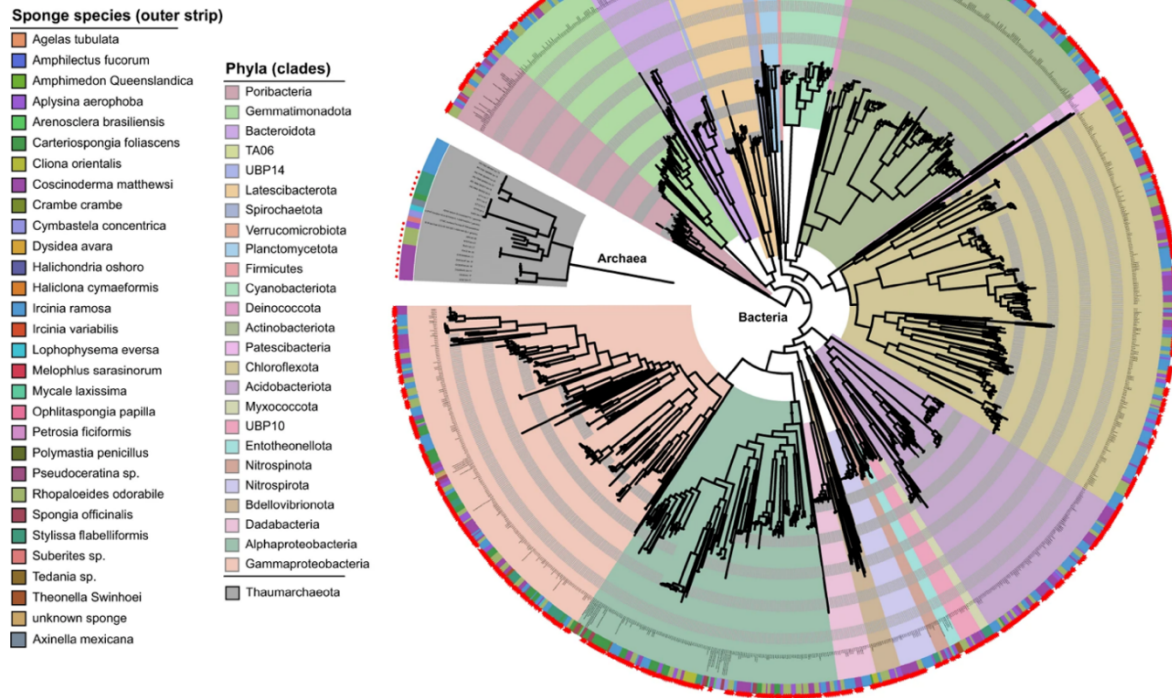


Figure 6. Distribution of microbial communities associated with coral reef demosponges, highlighting the taxonomic diversity of symbiotic microbes. Reproduced from Robbins et al., 2021.

2.5.3. Sponge Taxonomy

Sponge taxonomy integrates classical morphology with molecular approaches to resolve relationships within the phylum. Traditionally, sponges have been classified into four classes based on morphological characteristics such as skeletal composition, spicule shape, and overall body architecture. Demospongiae account for the vast majority of all described sponge species and represent the most diverse class in terms of morphology and habitat range. Members of this group typically possess siliceous spicule or spongin fibers. Demosponges occur in a wide variety of environments, including shallow marine ecosystems, deep-sea habitats, and freshwater systems. Calcareum is characterized by spicules composed of calcium carbonate. These sponges are generally small, tubular, or vase-shaped and are mostly restricted to shallow marine environments. Hexactinellida, commonly known as glass sponges, are distinguished by their siliceous spicules with a characteristic six-rayed structure. They possess a unique syncytial tissue organization, where large regions of the body consist of multinucleated continuous cytoplasm. Most species inhabit deep-sea environments, often forming extensive reef-like structures. Homoscleromorpha represents a relatively small but evolutionarily significant class of sponges. Members of this group possess a true basement membrane underlying their epithelial cells, as well as well-developed cell junctions, features more typical of bilaterian animals. Homoscleromorphs are typically found in marine

environments, often in cryptic habitats such as caves (Wörheide et al., 2012; Morrow & Cárdenas, 2015).

2.5.4. *Sponge Phylogeny*

Sponge phylogeny has been substantially revised through molecular and phylogenomic analyses, leading to a reassessment of early animal evolution. Most recent studies support the monophyly of Porifera and place sponges among the earliest-diverging metazoan lineages (Wörheide et al., 2012; Botting & Muir, 2018). However, their exact phylogenetic position relative to other basal animal groups remains uncertain. In particular, the identity of the sister group to all other animals is still debated: some analyses recover Porifera as the earliest-branching lineage (Redmond & McLysaght, 2021; Philippe et al., 2009), whereas others instead support ctenophores in this position (Moroz et al., 2014; Dunn et al., 2008).

Within the phylum, phylogenetic relationships derived from molecular data have clarified the distinctiveness of the four taxonomic classes. Demospongiae and Hexactinellida are often recovered as sister groups, reflecting their shared possession of siliceous spicules (Botting et al., 2017). In contrast, Calcarea forms a separate lineage (Manuel et al., 2003), while Homoscleromorpha is recognized as a distinct and early-branching clade based on both molecular evidence and morphological features (Gazave et al., 2010). Phylogenomic approaches using large-scale genomic and transcriptomic datasets continue to refine these relationships, highlighting patterns of gene gain, loss, and functional innovation associated with early metazoan evolution.

2.5.5. *Advances in Sponge Genomics*

Sponge genomics has advanced rapidly over the past decade, providing insights into the genetic foundations of early animal evolution. The first sponge genome to be sequenced was that of demosponge *A. queenslandica*, revealing a surprisingly rich repertoire of genes associated with cell signaling, adhesion, and developmental regulation. This demonstrated that key components of the metazoan genetic toolkit evolved before the emergence of complex body plans (Srivastava et al., 2010). Several subsequent genomes, including *Ephydatia muelleri* and *Tethya wilhelma*, have enabled comparative analyses within Demospongiae, confirming the conservation of major gene families alongside lineage-specific expansions, particularly in immunity and environmental response genes (Kenny et al., 2020; Francis et al., 2017). More recently, genomic resources have significantly expanded to other sponge classes, such as the glass sponge *Aphrocallistes vastus* (Francis et al., 2023) and the homoscleromorph *Corticium candelabrum* (Maldonado et al., 2025), allowing broader phylogenomic comparisons across the phylum.

Comparative genomic analyses reveal that sponges harbor surprisingly complex gene regulatory networks, including a diverse repertoire of transcription factor families that are also found in other metazoans. These include key developmental regulators such as homeobox, basic helix-loop-helix, Forkhead box, T-box, and nuclear receptor transcription

factors, many of which are associated with cell differentiation and various signaling pathways. This suggests that the foundations of animal developmental gene regulation were already established in early metazoans (Degnan et al., 2009; Larroux et al., 2006). At the same time, lineage-specific expansions and losses of transcription factor genes highlight the dynamic nature of sponge genome evolution, reflecting both the conservation of ancestral regulatory toolkits and the flexibility that has shaped their diversification (Kenny et al., 2020; Renard et al., 2018). Despite these advances, non-coding regions of sponge genomes remain poorly characterized. As noted earlier, the regulatory roles of lncRNAs are still largely unexplored, highlighting a significant gap in the understanding of gene regulation in early metazoans.

3. MATERIAL AND METHODS

3.1. *Methodological Overview*

This dissertation integrated the development of bioinformatic algorithms for transcriptome assembly quality assessment with a comparative analysis of lncRNAs across 14 sponge species. To facilitate lncRNA analysis, additional preprocessing steps were required, including reliable genome decontamination and comprehensive genome annotation. Accordingly, the workflow was structured into the following five components:

i) Development and benchmarking of two transcriptome quality assessment tools

Design and detailed performance evaluation of two novel algorithms for evaluating transcriptome assemblies.

ii) Retrieval and assembly of public sponge RNA-seq datasets

Systematic collection of publicly available RNA sequencing libraries representing diverse sponge taxa. Assembly of sponge transcriptomes, followed by extensive quality evaluation and improvement through integration of multiple assemblies.

iii) Genome decontamination pipeline development and application

Development and validation of a genome decontamination pipeline tailored for non-model organisms, followed by its application to the analysed sponge genomes to identify and remove contaminant sequences.

iv) Genomic annotation of TEs and PCGs

Annotation of sponge genomes, including the identification and characterization of TEs, followed by the prediction and functional annotation of PCGs.

v) Identification and annotation of lncRNAs

Characterization of lncRNAs across 14 sponge species.

This methodology ensured high-quality lncRNA identification and annotation through a comprehensive transcriptome quality assessment process, effective removal of contaminant sequences, and comprehensive genome annotation. This provided a robust foundation for detailed analyses of general lncRNA characteristics, TE associations, evolutionary conservation, and expression patterns.

3.2. CATS: Comprehensive Assessment of Transcript Sequences

This dissertation introduces Comprehensive Assessment of Transcript Sequences (CATS), a novel framework for transcriptome assembly evaluation that comprises two complementary components. CATS-rf performs reference-free quality assessment by evaluating the intrinsic accuracy of assembled transcripts. In contrast, CATS-rb provides reference-based evaluation by assessing structural accuracy and completeness of the assembly. This chapter is adapted from Bodulić & Vlahoviček (2026), in which CATS was originally developed and described.

3.2.1. CATS-rf Algorithm Overview

CATS-rf evaluates transcriptome assembly quality using short-read RNA-seq reads employed in the assembly process (Figure 7). The method aligns reads back to assembled transcripts and utilizes mapping evidence to generate an overall estimate of transcript accuracy.

The pipeline begins by mapping the RNA-seq library to the analysed assembly using Bowtie2 (Langmead & Salzberg, 2012), with parameters optimized for precise detection of assembly errors. Read alignment is performed in global configuration using the sensitive parameter preset. To allow for multi-mapping across similar transcripts, the default maximum number of alignments per read is set to 10. A mismatch penalty of 2 is applied to enable the detection of inaccurate transcript regions, while minimizing false positive mappings. Secondary alignments with edit distance exceeding a specific threshold are removed, reducing off-target mappings. This threshold is defined as the sum of the maximum edit distance of the read and 10% of read length.

Reads that align to multiple transcripts are resolved probabilistically, with assignment probability weighted using the transcripts per million (TPM) abundance estimate, ensuring that transcripts with higher expression levels are more likely to be assigned ambiguously mapped reads. These abundances are computed using Kallisto under default parameters (Bray et al., 2016). Specifically, the probability P of assigning a read to a given transcript t is defined as:

$$P = \frac{\sqrt{\text{TPM}_t}}{\sum_{t' \in S} \sqrt{\text{TPM}_{t'}}}$$

where S is the set of all candidate transcripts t' for a given read.

The introduced transformation of TPM aims to increase the effective number of assigned reads in transcripts with low expression levels. To ensure alignment consistency, CATS-rf enforces co-assignment of paired reads. If read pairs map to different transcripts, both are reassigned to the most abundant shared transcript. This eliminates artifacts that could bias downstream scoring components based on pair mapping.

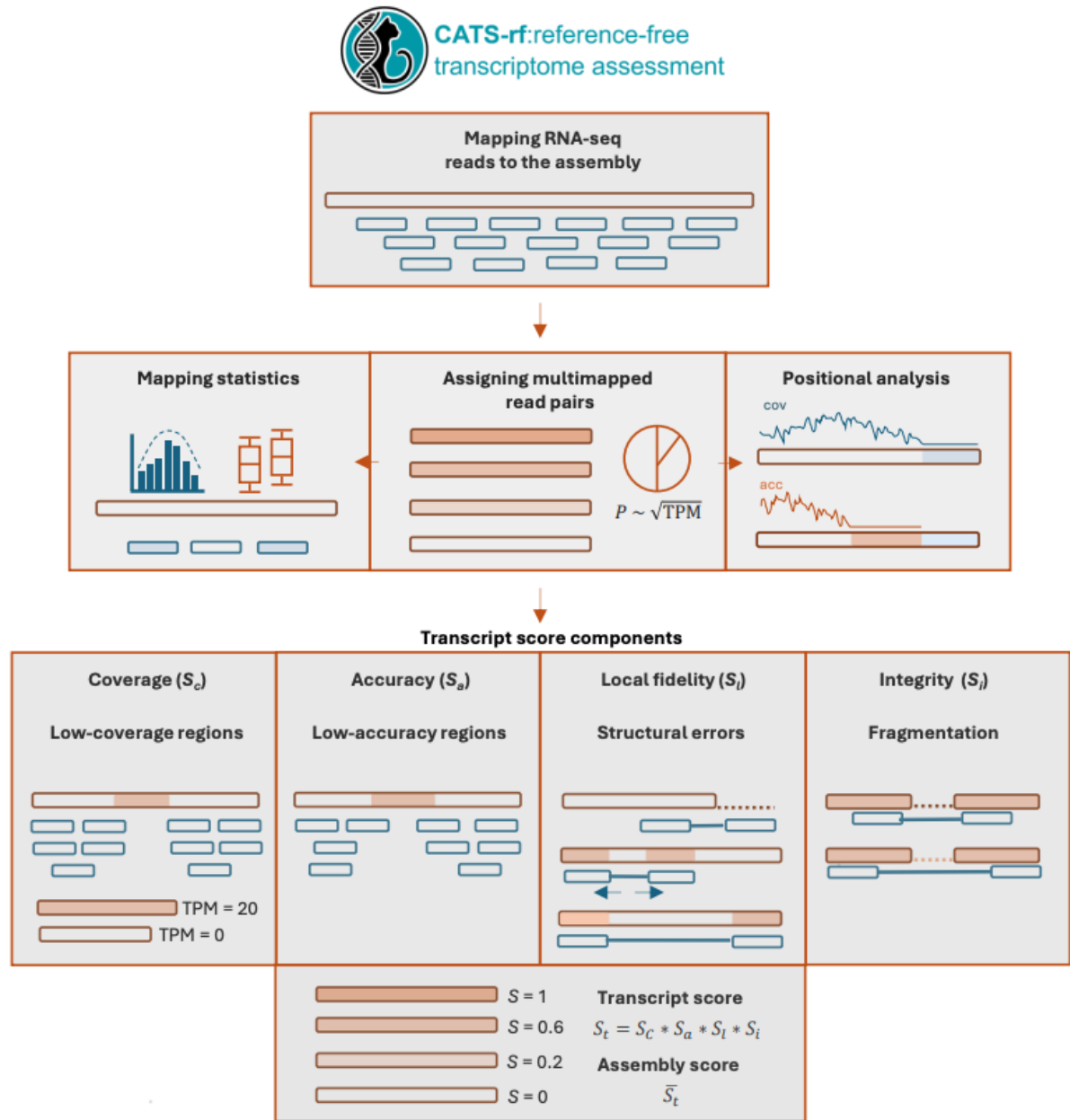


Figure 7. Overview of transcriptome assembly evaluation using CATS-rf. RNA-seq reads are aligned to assembled transcripts, with multimapped reads probabilistically assigned according to transcript abundance. Read alignments are used to compute four transcript-level score components: coverage, accuracy, local fidelity, and integrity, with each component capturing a distinct class of assembly errors. Transcript scores are defined as the product of the four components, with the assembly score computed as the mean of individual transcript scores. P = probability of read assignment, TPM = transcripts per million, S_c = coverage component, S_a = accuracy component, S_l = local fidelity component, S_i = integrity component, S_t = transcript score. Reproduced from Bodulić & Vlahoviček, 2026.

CATS-rf calculates transcript-level quality metrics by analyzing per-base coverage and accuracy patterns, along with the consistency of paired-read mappings. Per-base statistics are extracted from read alignments using pysamstats (Miles, 2013). Four score components are defined, each targeting distinct types of assembly errors.

The coverage score component S_c penalizes transcript regions with insufficient read support, primarily aiming to detect unsupported insertions and transcript redundancy. Low coverage regions (LCRs) are detected by calculating the mean per-base coverage within rolling windows of fixed length and merging all contiguous bases whose corresponding window-mean coverage falls below a predefined threshold. Each LCR is assigned a coverage penalty P_c computed as:

$$P_c = \sum_{n=1}^{L_c} f(C_n) + E_c * L_c$$

Here, C_n represents the coverage of the n th base. E_c denotes the LCR extension penalty, L_c corresponds to LCR length, and $f(C_n)$ is a function of per-base coverage defined as:

$$f(C_n) = \begin{cases} 1, & \text{if } C_n = 0 \\ \frac{1}{W * C_n}, & \text{if } C_n > 0 \end{cases}$$

where W denotes the base coverage weight, which scales the penalty applied to bases with non-zero coverage.

The coverage penalty incorporates per-base coverage ($f(C_n)$) and an extension term (E_c), capturing the magnitude and extent of reduced coverage within each LCR. This definition enables the detection of extended unsupported regions indicative of assembly errors. In contrast, read- or allele-specific indels incur minimal penalties, which is appropriate considering that such events represent false-positive coverage drops rather than genuine misassemblies.

The coverage component is calculated as the complement of the normalized sum of coverage penalties:

$$S_c = 1 - \sqrt{\frac{\sum_{n=1}^{N_c} P_{c,n}}{L_t * (E_c + 1)}}$$

where N_c is the total number of LCRs per transcript and L_t corresponds to transcript length.

The coverage component normalizes the total coverage penalty per transcript such that a value of zero indicates a complete lack of coverage, while a value of one is assigned to transcripts without detected LCRs.

The accuracy score component S_a detects regions with inaccurate transcript sequence. In this context, accuracy is defined as the proportion of aligned read bases matching the corresponding transcript base. Low-accuracy regions (LARs) are defined by calculating the mean accuracy of covered bases within rolling windows of fixed length and merging contiguous bases whose window-mean accuracy falls below a defined threshold. This definition is further refined by excluding fully accurate bases on LAR ends. Each LAR is assigned an accuracy penalty P_a given by:

$$P_a = \sum_{n=1}^{L_a} (1 - A_n) + E_a * L_a$$

where A_n is the accuracy of the n_{th} base, E_a represents the LAR extension penalty, and L_a corresponds to LAR length.

Analogous to its coverage counterpart, the accuracy penalty combines per-base accuracy with an extension term to capture both the degree and extent of reduced accuracy. This formulation emphasizes longer inaccurate regions, which are indicative of misassembly events such as the collapse of similar isoforms, transcripts originating from paralogous gene families, or repetitive sequences. In contrast, isolated base-level discrepancies, including sequencing errors or single nucleotide polymorphisms (SNPs), incur only minor penalties, which is appropriate as they do not represent true assembly errors.

The accuracy component is calculated as the complement of the normalized sum of accuracy penalties:

$$S_a = 1 - \sqrt{\frac{\sum_{n=1}^{N_a} P_{a,n}}{L_{cov} * (E_a + 1)}}$$

where N_a is the number of LARs per transcript and L_{cov} represents the total number of covered transcript bases.

The accuracy component normalizes the total accuracy penalty by the number of covered bases per transcript such that a value of zero represents the theoretical scenario in which all covered bases have zero accuracy, and a value of one is assigned to transcripts without detected LARs.

When a paired-end library is supplied, CATS-rf computes the local fidelity and integrity score components using paired-read mapping patterns. The local fidelity score component S_f

identifies mapping inconsistencies of read pairs within individual transcripts, flagging structural errors such as deletions, inversions, or translocations. This component integrates three transcript-level metrics: number of reads whose pair fails to map to the assembly (N_u), the number of read pairs mapping to the same transcript in unexpected orientations (N_o), and the total distance penalty (P_d) per transcript. Unexpected orientations are determined based on library strandness, which can be supplied by the user or inferred automatically.

The distance penalty P_d is calculated for each properly oriented pair mapped to the same transcript using a segmented function of the observed read pair distance d :

$$P_d = \begin{cases} 0, & \text{if } d < D_1 \\ \frac{d}{D_2}, & \text{if } D_1 \leq d \leq D_2 \\ 1, & \text{if } d > D_2 \end{cases}$$

where D_1 and D_2 represent distance outlier thresholds calculated as:

$$D_1 = Q_3(d) + M_1 * (IQR(d) + C)$$

$$D_2 = Q_3(d) + M_2 * (IQR(d) + C)$$

Here, $Q_3(d)$ denotes the third quartile and $IQR(d)$ is the interquartile range of the observed pair distance distribution. The multiplicative factors M_1 and M_2 reduce the influence of false-positive outliers that may arise from variation in library preparation. The correction factor C further stabilizes threshold robustness in libraries with a high proportion of overlapping read pairs. This formulation penalizes read pairs exhibiting distance outliers relative to the expected interpair distance distribution, where the outlier threshold is determined by D_1 , and the penalty scaling is controlled by D_2 . By definition, the distance penalty is normalized to the interval $[0, 1]$,

The described mapping inconsistencies are combined into the local fidelity component:

$$S_l = 1 - \sqrt{\frac{N_u + 2N_o + 2 \sum_{n=1}^{N_p} P_{d,n}}{N_r}}$$

where N_p denotes the number of properly oriented pairs mapped to the same transcript and N_r represents the total number of mapped reads per transcript. The local fidelity component quantifies the consistency of paired-read mapping by normalizing the number of improperly mapped reads within a transcript by the total number of mapped reads, yielding values in the interval $[0, 1]$.

The integrity score component S_i captures transcript fragmentation by identifying bridging events, defined as read pairs that map to different transcripts. For each bridging event, a

bridge penalty P_b is calculated as the mean of the relative distances of each read from the center of its respective transcript. These distances are scaled such that a value of 0 corresponds to the transcript center and a value of 1 to the transcript ends. The definition assumes that bridging read pairs located near transcript ends are more likely to reflect true fragmentation. The bridge index B for a transcript is then computed by normalizing the sum of bridge penalties against the total number of mapped reads per transcript:

$$B = 1 - \frac{1}{N_r} \sum_{n=1}^{n_b} P_{b,n}$$

where n_b denotes the number of bridging events per transcript.

The integrity component is computed by applying a sigmoid transformation to the bridge index:

$$S_l = \frac{B^\alpha}{B^\alpha + (1 - B)^\beta}$$

Here, α and β represent scaling parameters that regulate sensitivity to fragmentation patterns. The sigmoid transformation accounts for the consistent mapping of read pairs to accurately assembled interior transcript regions, which inherently weakens the fragmentation signal.

The transcript score S_t is calculated as the product of the described score components:

$$S_t = S_c * S_a * S_l * S_i$$

The assembly score S is computed as the mean of individual transcript scores:

$$S = \frac{1}{N_t} \sum_{n=1}^{N_t} S_{t,n}$$

where N_t denotes the total number of transcripts. The assembly score provides an overall estimate of assembly quality by equally weighting distinct classes of assembly errors. Both transcript and assembly scores are normalized to the interval $[0, 1]$, with higher values indicating better quality. Transcripts lacking any read coverage are assigned a transcript score of zero, as their coverage component is set to zero, while the remaining components are left undefined.

Because the assembly score captures only the accuracy of the reconstructed transcripts, CATS-rf additionally reports assembly completeness as a separate metric. This measure reflects the

proportion of input reads that successfully align to the assembled transcripts, providing an estimate of how comprehensively the assembly represents the underlying RNA-seq data.

Additionally, CATS-rf generates and visualizes a broad suite of mapping and positional metrics to enable detailed quality assessment. Coverage and accuracy statistics include the total number of uncovered and inaccurate bases, as well as summary metrics of per-base coverage and accuracy. The tool further provides positional distributions of coverage and accuracy, enabling the identification of systematic biases along transcripts. CATS-rf also offers an in-depth characterization of LCRs and LARs, delivering detailed insights into problematic transcript segments. Furthermore, it generates a comprehensive report of paired-end read mapping metrics, including a detailed characterization of improperly paired reads, bridging events, and transcript fragmentation patterns.

Considering that CATS-rf score components include several tunable parameters, detailed descriptions of their default values and the underlying rationale are provided in Table 1. It should be noted that none of these parameters were specifically optimized for the evaluated assemblies. Instead, default parameter values were chosen to reflect moderate biological characteristics and library preparation strategies.

Table 1. Default values of tunable parameters included in CATS-rf score component calculation. Adapted from Bodulić & Vlahoviček, 2026.

Score component	Feature	Parameter	Default value	Rationale
Coverage (S_c)	Low-coverage region (LCR)	Rolling window length for LCR detection (bp)	10	The default length and threshold values were selected to differentiate regions with uniformly low read support from isolated single-base coverage gaps caused by read indels or transcript variation
	Low-coverage region (LCR)	Mean rolling window coverage threshold for LCR definition	3	
	Coverage penalty (P_c)	LCR Extension penalty (E_c)	0.5	The default value of E_c was selected to impose a moderate penalty for increased LCR length relative to per-base coverage. Extending the LCR by one base increases the coverage penalty equivalently to adding two bases with zero coverage
	Per-base coverage function ($f(C_n)$)	Base coverage weight (W)	1.5	The default value of W moderately lowers the penalty assigned to LCR bases with non-zero coverage
Accuracy (S_a)	Low-accuracy region (LAR)	Rolling window length for LAR detection (bp)	10	The default length and threshold values were selected to identify extended regions of lower accuracy supporting misassembly, while reducing the penalization of sequencing errors occurring at standard Illumina error rates
	Low-accuracy region (LAR)	Mean rolling window accuracy threshold for LAR definition	0.98	
	Accuracy penalty (P_a)	LAR Extension penalty (E_a)	0.1	The default value of E_a was selected to incur a moderate penalty for increased LAR length relative to per-base accuracy. Extending the LAR by one base increases the total accuracy penalty equivalently to adding five bases with 0.5 accuracy

Local fidelity (S_l)	Lower distance threshold (D_1)	Lower multiplicative factor (M_1)	8	The default value of M_1 was chosen to minimize false-positive penalties resulting from library-specific variation in fragment length distributions
	Higher distance threshold (D_2)	Higher multiplicative factor (M_2)	10	The default value of M_2 imposes a 2-IQR extension relative to D_1 , such that outliers between 8 and 10 IQRs beyond Q_3 are scaled from 0 to 1, while stronger outliers are maximally penalized
	Lower and higher distance thresholds (D_1 and D_2)	Correction factor (C)	5	The default value of C was selected to provide appropriate distance penalization in libraries with high frequency of pair overlaps, while ensuring low distance penalty inflation
Integrity (S_i)	Integrity component (S_i)	Compression factor 1 (α)	7	The default values of α and β were selected to produce a balanced sigmoid curve that is sufficiently steep to penalize fragmentation while remaining stable for intact transcripts
	Integrity component (S_i)	Compression factor 2 (β)	0.5	

bp = base pairs

3.2.2. CATS-rb Algorithm Overview

CATS-rb evaluates transcriptome assemblies by aligning transcripts to the reference genome of the corresponding species (Figure 8). The pipeline comprises three core tools that perform genome index generation, transcriptome mapping, and comparative mapping analysis. It calculates a wide range of structural accuracy metrics and quantifies transcriptome completeness at both transcript and exon levels. CATS-rb is designed to analyse multiple transcriptomes simultaneously, enabling direct comparison of evaluated assemblies.

CATS-rb uses Spaln to map transcripts to a reference genome, leveraging its high accuracy for spliced alignments. Spaln employs sophisticated dynamic programming algorithms and species-aware scoring schemes to accurately map transcripts even in the presence of sequencing errors and structural variation (Gotoh, 2024).

Transcriptome assembly mappings are compared in terms of structural accuracy and completeness. The analysis begins by processing exon-level alignment output to reconstruct transcript-level mappings. Exons derived from the same transcript are assigned to separate transcript mappings when any of the following conditions are met: (i) the exon overlaps the preceding exon in transcript coordinates by more than half of the preceding exon's length; (ii) the exon maps to a different genomic scaffold or strand than the preceding exon; or (iii) the exon maps to the same scaffold and strand as the preceding exon but is separated by a distance exceeding the defined intron length threshold.

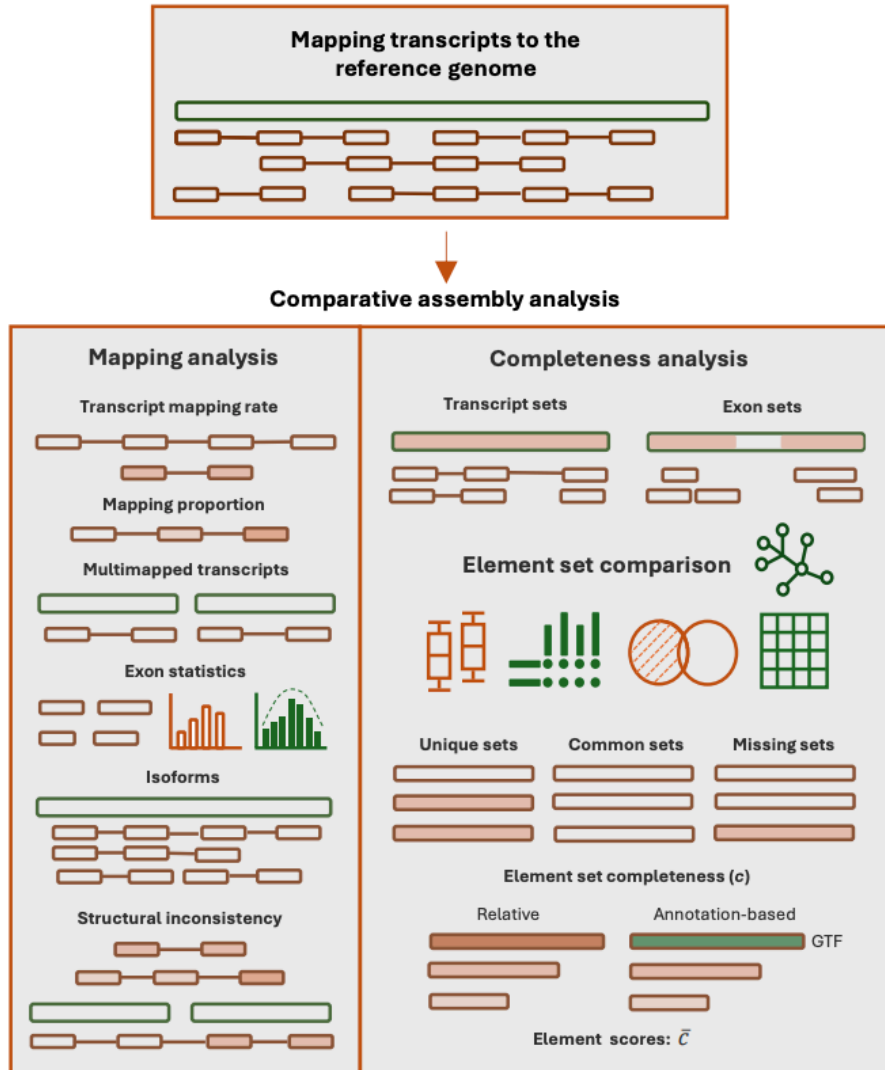


Figure 8. Overview of transcriptome assembly evaluation by CATS-rb. Transcripts are mapped to the reference genome and several mapping metrics are evaluated, including mappability, exon-level statistics, and structural consistency. Completeness analysis is performed across two categories of element sets. Specifically, transcript and exon sets are determined by overlapping transcript or exon genomic coordinates within each transcriptome. Element set coordinates are overlapped between transcriptomes, constructing a graph where edges represent element set overlaps and connected components denote element set groups. In-depth analysis of element sets is performed, visualized using UpSet plots, Venn diagrams, and hierarchical clustering heatmaps. Element set completeness is defined as the relative length of each set compared to the longest set within its group. The analysis can be extended using reference element sets derived from genomic annotation. Transcript and exon scores are calculated as the mean completeness of the respective element sets. Reproduced from Bodulić & Vlahoviček, 2026.

CATS-rb reports multiple transcriptome assembly mapping metrics, including transcript mappability and multimapping, exon count, exon length distribution, isoform representation, and structural consistency. A transcript is considered structurally inconsistent if its alignment proportion falls below a specific threshold or if it contains regions that map to disjunct genomic loci. The latter is evaluated by identifying transcript segments that overlap by less than a predefined fraction of their lengths and map either to different scaffolds, to opposite strands on the same scaffold, or to distant regions on the same scaffold and strand (beyond the specified intron length threshold). Flagged candidates are excluded if one segment is shorter than 10% of the other segment's length. The definition is further refined by removing cases where a larger, separately mapped transcript segment overlaps with more than 90% of each segment in the flagged pair (in transcript coordinates). The structural consistency criteria are primarily intended to capture chimeric transcripts, while also identifying substantial mismatches to the reference genome arising from misassembly or genuine biological variation.

The main contribution of CATS-rb is the comparative analysis of transcriptome assembly completeness, which introduces two types of element sets. Precisely, transcript and exon sets are defined by merging the overlapping transcript or exon genomic coordinates within each transcriptome into non-redundant segments using GenomicRanges (Lawrence et al., 2013). For multimapped transcripts, element sets are built using the most accurate alignment, with ties resolved by choosing the matching alignment across the analysed transcriptomes. CATS-rb constructs separate undirected graphs for both element set types using igraph (Csárdi et al., 2026). In these graphs, nodes represent element sets and edges indicate their overlaps between the compared transcriptomes. Only overlaps exceeding a predefined length threshold are retained. The resulting transcript and exon set graphs are partitioned into connected components to define transcript and exon set groups, respectively. If multiple sets from the same transcriptome are contained in a single group, only the longest set is retained. The group representative is then determined as the longest set in the group.

Element set completeness c is defined as the length of the set relative to the representative set of its group. Sets absent from a group are assigned a completeness of 0. The relative element score R for each transcriptome is calculated as the mean element set completeness across all groups:

$$R = \frac{1}{N_g} \sum_{n=1}^{N_g} c_n$$

Here, N_g represents the total number of element set groups. This approach is applied independently to transcript and exon sets to produce relative transcript and exon scores.

The completeness analysis also offers a comprehensive evaluation of element set distribution across the compared transcriptome assemblies. This includes a detailed analysis of unique,

common, and missing element sets, along with element set UpSet plots and Venn diagrams. Transcriptomes are also hierarchically clustered according to element set completeness using complete linkage clustering with the Euclidean distance metric.

The presented algorithm evaluates transcriptome assembly completeness relative to the compared transcriptomes without relying on external genomic annotation. However, CATS-rb can also perform an annotation-based analysis utilizing reference element sets. These sets are formed by separately merging overlapping transcript and exon genomic coordinates defined by genomic annotation. The annotation-based approach follows the same methodology as relative completeness evaluation, while grouping transcriptome element sets based on shared overlaps with reference sets. Accordingly, reference sets are considered the representative for each element set group. Annotation-based element scores A are calculated analogously to relative element scores, providing an absolute metric of transcriptome completeness.

Additionally, genomic annotation can be used to compute element set sensitivity and specificity for each transcriptome assembly. Sensitivity is calculated as the proportion of reference sets matched by transcriptome element sets, while specificity reflects the proportion of transcriptome element sets matched by reference sets. A set is considered matched if its genomic coordinates overlap with more than a predefined fraction of the corresponding set's length.

CATS-rb offers a broad set of assembly-specific parameters that allow analyses to be adapted to the biological and technical characteristics of the dataset. In the mapping stage, parameters control transcript alignment behavior, including strandedness, the maximum number of alignments per transcript, and the use of species-specific presets. Splice-site recognition stringency can also be adjusted, and alignment scoring can be tuned by weighting contributions from coding potential and translational signals, thus enabling prioritization of protein-coding transcripts (PCTs). In the comparative analysis stage, CATS-rb allows control over several thresholds, with the most relevant parameters including minimum exon identity, exon and intron length thresholds, and element set length thresholds. Permissive configurations enhance sensitivity to features with low exon identity, short exons, and small element sets, but may also increase susceptibility to spurious signals. Conversely, more stringent thresholds reduce noise while potentially excluding genuine but shorter features. In general, more sensitive parameter settings are best suited to less complex genomes with low risk of spurious signal detection.

Table 2. Default values of tunable parameters included in CATS-rb methodology.

Stage	Parameter	Default value
Genome index generation	Maximum gene size (bp)	Estimated from genome size
Transcriptome mapping	Strand-aware mapping	Off
	Maximum number of alignments per transcript	5
	Minimum intron length (bp)	20
	Species-specific mapping preset	Undefined
	Allowed mismatches from canonical splice sites	1
	Coding potential weight for alignment scoring	1
	Translation initiation signal weight for alignment scoring	1
Comparative mapping analysis	Minimum exon identity proportion	0.98
	Minimum exon length (bp)	20
	Maximum intron length (bp)	100000
	Minimum alignment proportion for structural inconsistency detection	0.9
	Minimum overlap for isoform specification (bp)	1
	Minimum exon set length (bp)	0
	Minimum transcript set length (bp)	100
	Maximum transcript set length (bp)	1000000
	Minimum overlap between exon sets for edge specification (bp)	1
	Minimum overlap between transcript sets for edge specification (bp)	1

bp = base pairs

3.2.3. CATS Benchmarking Library Preparation

A detailed assessment of CATS performance was conducted using a broad range of simulated and public RNA-seq libraries. Simulation was performed using the Polyester R package (R version 4.4.2, package version 1.32.0) (R Core Team, 2026; Frazee et al., 2015). The initial simulation employed a controlled framework in which transcriptome complexity, transcript coverage ranges, sequencing error rate, and assembler choice were varied orthogonally. This encompassed a total of 126 unstranded paired-end libraries representing coding and non-coding transcripts of six species: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Mus musculus*, and *Homo sapiens* (Table 3). For each species, 21 libraries were constructed by combining seven transcript coverage levels with three sequencing error rates. Coverage levels were defined across the following ranges: 1–4×, 5–10×, 11–20×, 21–30×, 31–40×, 41–50×, and 51–100×. Within each library, transcripts were randomly assigned a coverage value drawn from the corresponding range, with all transcripts considered expressed. A uniform error model was applied, with sequencing error rates set to 0.005, 0.01, or 0.02 for each coverage level.

Table 3. Reference transcriptome assemblies used in simulating RNA-seq reads for CATS benchmarking. Reproduced from Bodulić & Vlahoviček, 2026.

Species	Build	Ensembl release	N transcripts	Total length (Mb)
<i>S. cerevisiae</i>	R64-1-1	111	7036	8.9
<i>C. elegans</i>	WBcel235	111	60000	58.2
<i>D. melanogaster</i>	BDGP6.46	111	35722	92.8
<i>A. thaliana</i>	TAIR10	56	54013	89.2
<i>M. musculus</i>	GRCm39	111	145855	243.9
<i>H. sapiens</i>	GRCh38	111	275921	471.4

Mb = megabase

CATS benchmark was also performed on realistically simulated libraries with multifold variation in transcript coverage. A total of 96 libraries were generated, spanning the same six species and four different coverage levels, each represented by four replicates (Table 4, Table 5). A uniform base-level sequencing error rate of 0.005 was applied. Library coverage was defined by a target mean coverage level C chosen from four possible values: 20×, 40×, 60×, and 80×. The number of read pairs per library N_r was computed as:

$$N_r = \frac{C * \sum_{n=1}^{N_e} L_{t,n}}{2 * L_r}$$

where N_e is the number of expressed transcripts, $L_{t,n}$ represents the length of the n th transcript, and L_r denotes the read length.

To reflect realistic biological variability, the fraction of expressed transcripts per replicate was randomly sampled from species-specific ranges: *S. cerevisiae*: 70–90%, *C. elegans*, *D. melanogaster*, and *A. thaliana*: 30–50%, *M. musculus* and *H. sapiens*: 20–40%. For each expressed transcript, relative abundance values were sampled from a log-normal distribution with parameters meanlog = 1 and sdlog = 2, corresponding to a $\sim 10^4$ – 10^5 -fold range in coverage within a single library (Figure 9). These values were subsequently scaled by transcript length. The number of reads originating from each transcript was calculated by multiplying relative abundance values by the total number of reads per library. In both controlled and realistic simulations, reads were 100 bp in length, with fragments generated using an RNA fragmentation protocol based on the empirical fragment length distribution provided by Polyester. Per-transcript coverage values were plotted using R (version 4.4.2) using the ggplot2 package (version 3.5.1) (Wickham, 2016).

Table 4. Number of expressed transcripts in realistically simulated RNA-seq libraries used for CATS benchmarking. Reproduced from Bodulić & Vlahoviček, 2026.

Species	Replicate	N (%) expressed transcripts
<i>S. cerevisiae</i>	1	5136 (73%)
	2	5558 (79%)
	3	5981 (85%)
	4	5558 (79%)
<i>C. elegans</i>	1	30000 (50%)
	2	21000 (35%)
	3	30000 (50%)
	4	25200 (42%)
<i>D. melanogaster</i>	1	12146 (34%)
	2	15361 (43%)
	3	15718 (44%)
	4	13932 (39%)
<i>A. thaliana</i>	1	21605 (40%)
	2	18364 (34%)
	3	21605 (40%)
	4	23766 (44%)
<i>M. musculus</i>	1	30630 (21%)
	2	37922 (26%)
	3	39381 (27%)
	4	49591 (34%)
<i>H. sapiens</i>	1	110368 (40%)
	2	110368 (40%)
	3	99312 (36%)
	4	63462 (23%)

Table 5. Number of reads in realistically simulated RNA-seq libraries used for CATS benchmarking. Reproduced from Bodulić & Vlahoviček, 2026.

Species	Coverage	N reads (million)			
		Replicate 1	Replicate 2	Replicate 3	Replicate 4
<i>S. cerevisiae</i>	20×	0.6	0.7	0.7	0.7
	40×	1.3	1.4	1.5	1.4
	60×	1.9	2.1	2.2	2.1
	80×	2.6	2.8	3.0	2.8
<i>C. elegans</i>	20×	2.9	2.0	3.0	2.5
	40×	5.8	4.0	5.9	4.9
	60×	8.7	6.0	8.8	7.4
	80×	11.6	8.0	11.8	9.8
<i>D. melanogaster</i>	20×	3.2	4.0	4.1	3.6
	40×	6.4	8.0	8.2	7.2
	60×	9.6	11.9	12.3	10.8
	80×	12.8	15.9	16.5	14.4
<i>A. thaliana</i>	20×	3.6	3.0	3.6	4.0
	40×	7.1	6.1	7.1	7.9
	60×	10.7	9.1	10.7	11.8
	80×	14.3	12.2	14.3	15.8
<i>M. musculus</i>	20×	5.1	6.3	6.6	8.3
	40×	10.2	12.6	13.2	16.5
	60×	15.3	18.9	19.8	24.8

	80×	20.5	25.2	26.4	33.0
<i>H. sapiens</i>	20×	18.7	18.9	17.0	10.8
	40×	37.4	37.8	34.0	21.7
	60×	56.1	56.6	50.9	32.5
	80×	74.9	75.4	68.0	43.3

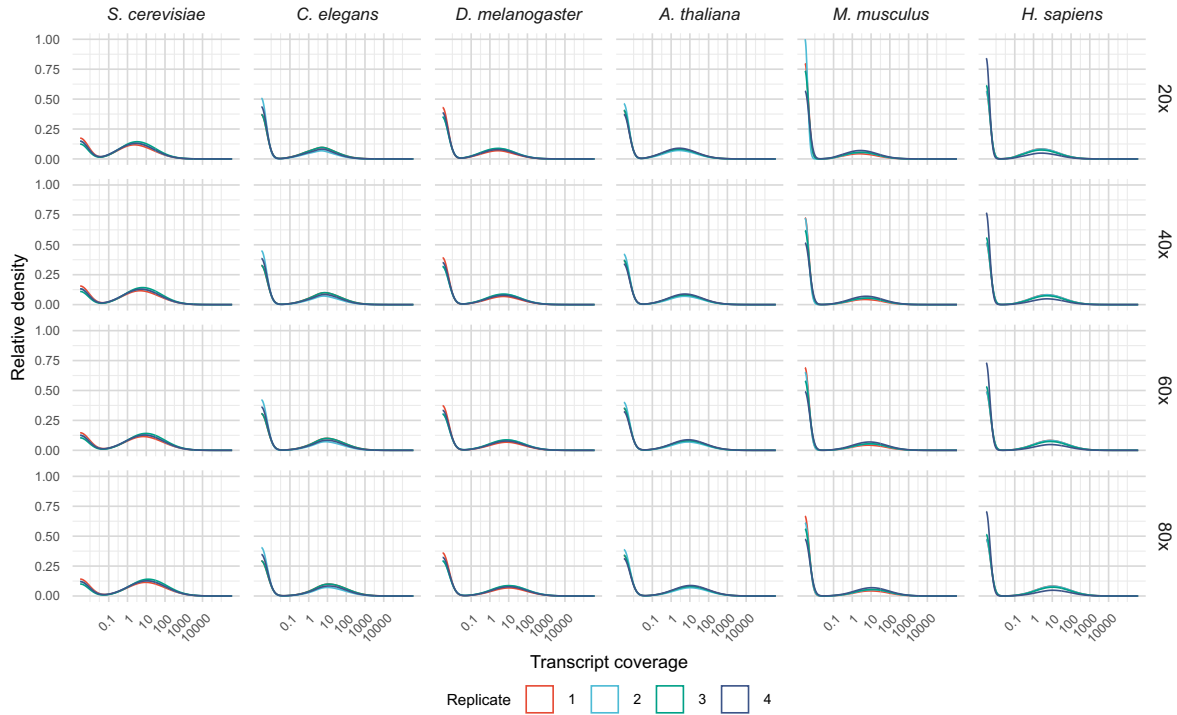


Figure 9. Reference transcript coverage distribution in realistically simulated datasets. Reproduced from Bodulić & Vlahoviček, 2026.

CATS was also benchmarked on 42 public RNA-seq libraries from the same six species utilized in controlled and realistic simulations (Table 6). Public libraries were randomly selected from the Sequence Read Archive (SRA) database with the SRA Toolkit (version 3.2.0) (Leinonen et al., 2011). Adapter sequences were removed and low-quality bases trimmed using Trim Galore (version 0.6.1) with cutadapt (version 4.6) (Martin, 2011). Adapter trimming required a minimum overlap of five nucleotides between the read and adapter sequence. Quality trimming involved the removal of terminal bases with a Phred quality score below 5 (parameters --stringency 5 -q 5).

Table 6. Public RNA-seq libraries used in CATS benchmarking. Reproduced from Bodulić & Vlahoviček, 2026.

SRA accession	Species	Publication date (DD.MM.YYYY)	Mean read length (bp)	N read pairs (million)	Total size (Gb)
SRR32108057	<i>S. cerevisiae</i>	24.01.2025	150	27.2	8.2
SRR30685385	<i>S. cerevisiae</i>	17.09.2024	72	2.2	0.3
SRR27822254	<i>S. cerevisiae</i>	26.03.2024	51	17.9	1.8
SRR26147123	<i>S. cerevisiae</i>	31.12.2023	146	7.0	2.0
SRR26001199	<i>S. cerevisiae</i>	11.11.2023	74	5.5	0.8
SRR24356101	<i>S. cerevisiae</i>	29.04.2023	150	19.8	1.8
SRR10815431	<i>S. cerevisiae</i>	21.02.2020	150	3.7	0.4
SRR32732688	<i>C. elegans</i>	17.03.2025	151	31.1	9.4
SRR32058534	<i>C. elegans</i>	21.01.2025	150	27.6	8.3
SRR31107386	<i>C. elegans</i>	20.01.2025	150	26.4	7.9
SRR12868377	<i>C. elegans</i>	10.02.2022	150	60.7	18.2
SRR17736005	<i>C. elegans</i>	25.01.2022	50	35.1	3.6
SRR5224028	<i>C. elegans</i>	01.03.2018	100	36.2	7.2
SRR6266308	<i>C. elegans</i>	25.02.2018	125	27.1	6.8
SRR31530981	<i>D. melanogaster</i>	28.11.2024	101	28.6	5.8
SRR29606991	<i>D. melanogaster</i>	28.06.2024	151	31.9	9.6
SRR24876155	<i>D. melanogaster</i>	17.11.2023	151	30.4	10.4
SRR23933588	<i>D. melanogaster</i>	27.06.2023	51	87.8	9.0
SRR23870057	<i>D. melanogaster</i>	17.03.2023	80	45.3	7.2
SRR20326862	<i>D. melanogaster</i>	25.07.2022	150	32.9	9.9
SRR14560308	<i>D. melanogaster</i>	01.01.2022	150	19.7	5.8
SRR33339778	<i>A. thaliana</i>	30.04.2025	150	37.6	11.3
SRR30855151	<i>A. thaliana</i>	11.10.2024	147	27.9	8.2
SRR14160649	<i>A. thaliana</i>	12.11.2023	101	37.5	7.6
SRR24948904	<i>A. thaliana</i>	20.10.2023	150	44.0	13.2
SRR24636232	<i>A. thaliana</i>	08.10.2023	150	31.0	9.3
SRR13159213	<i>A. thaliana</i>	31.12.2021	101	43.7	8.8
SRR8422221	<i>A. thaliana</i>	23.01.2019	74	33.0	4.9
SRR32903297	<i>M. musculus</i>	28.03.2025	150	28.3	8.5
SRR31588239	<i>M. musculus</i>	03.12.2024	50	41.0	4.1
SRR27369179	<i>M. musculus</i>	11.07.2024	151	66.2	20.0
SRR28908270	<i>M. musculus</i>	07.06.2024	101	62.9	12.7
SRR18855849	<i>M. musculus</i>	01.10.2023	151	49.9	15.1
SRR24134730	<i>M. musculus</i>	15.08.2023	100	7.4	3.5
SRR13981556	<i>M. musculus</i>	15.10.2021	151	28.3	8.5
SRR32139880	<i>H. sapiens</i>	27.01.2025	150	25.9	7.8
SRR31785722	<i>H. sapiens</i>	30.12.2024	150	71.8	21.5
SRR25646397	<i>H. sapiens</i>	13.06.2024	150	45.7	13.7
SRR25732618	<i>H. sapiens</i>	13.12.2023	101	50.5	10.2
SRR22548603	<i>H. sapiens</i>	21.12.2022	101	53.8	10.9
SRR8357441	<i>H. sapiens</i>	02.06.2019	100	51.7	10.3
SRR7741229	<i>H. sapiens</i>	26.03.2019	51	40.5	4.1

SRA = Short Read Archive, bp = base pairs, Gb = gigabase

Transcriptomes from simulated and public libraries were assembled using four relevant de novo assemblers: rnaSPAdes (version 3.15.4) (Bushmanova et al., 2019), Trinity (version 2.15.1) (Grabherr et al., 2011), IDBA-tran (version 1.1.1) (Peng et al., 2013), and SOAPdenovo-Trans (version 1.0.4) (Xie et al., 2014), each with default parameters. This generated a total of

504 assemblies from controlled simulations, 384 assemblies from realistic simulations, and 168 assemblies from public libraries.

Several aspects of CATS performance were also evaluated on artificially mutated assemblies. Firstly, mutations were independently introduced into a subset of assemblies from controlled simulations, including insertions, mismatches, deletions, fragmentation, and redundancy. Mutations were simulated in rnaSPAdes and Trinity assemblies from libraries of the six previously described species, each with transcript coverage of 21–30× and a sequencing error rate of 0.005 ($N = 12$). For every mutation type, four severity levels were simulated. In the case of insertions, mismatches, and deletions, a single mutated region was introduced at a random position within the sampled transcripts. The proportion of affected transcripts was set to 15%, 30%, 45%, and 60%, with the length of the mutated regions scaled accordingly (15%, 30%, 45%, or 60% of the transcript length). For mismatches, each base within the affected region was given a 25% probability of substitution, with equal likelihood across alternative bases. Fragmentation was simulated by splitting transcripts at a random position between the 40th and 60th percentiles of transcript length, affecting 20%, 40%, 60%, and 80% of transcripts, respectively. Redundancy was introduced by duplicating transcript segments beginning from the 5' transcript end, with 20%, 40%, 60%, and 80% of transcripts containing duplicated regions covering 20–39%, 40–59%, 60–79%, and 80–99% of the transcript length, respectively.

Additionally, multiplicative error simulation was performed, sequentially applying all five mutation types to the same set of 12 transcriptome assemblies. The mutations were introduced in the following order: deletion, mismatch, insertion, fragmentation, and redundancy. The simulation was conducted at six mutation severity levels (10%, 20%, 30%, 40%, 50%, and 60%), controlling both the proportion of mutated transcripts and, where applicable, the length of the affected region. To introduce variation, the multiplicative simulation was performed in triplicates. All mutation simulations were performed using R (version 4.4.2) and R package Biostrings (version 2.27.1) (Pagès et al., 2025).

3.2.4. CATS-rf Benchmarking

To compare the performance of CATS-rf, RSEM-EVAL (version 1.11) (Li et al., 2014), and TransRate (version 1.0.4) (Smith-Unna et al., 2016), all simulated and public transcriptome assemblies were analysed with each tool using default parameter settings. For simulated assemblies, the fragment length distribution mean required by RSEM-EVAL was set to 180 bp, consistent with the empirical distribution provided by Polyester (Frazee et al., 2015). For public assemblies, fragment length estimates were calculated using Kallisto (version 0.50.1). TransRate was downloaded from a GitHub repository that addressed inaccuracies produced by the latest official release (Madeira, 2016). To quantify true transcript-level accuracy, CRB-BLAST (version 1.0.0) (Aubry et al., 2014) was used to identify reciprocal best alignments between each tested transcript and reference transcripts. Transcript accuracy was quantified using the F-score, defined as the harmonic mean of transcript precision and recall. Here,

precision was defined as the proportion of aligned bases relative to the tested transcript length, while recall was computed as the proportion of aligned bases relative to the reference transcript length. Tested transcripts without CRB-BLAST hits were assigned an F-score of 0. For analyses of realistically simulated libraries, CRB-BLAST was performed using only the reference transcripts present in each replicate, as defined by the simulation design. For assemblies generated from public RNA-seq libraries, CRB-BLAST was restricted to evaluated transcripts that were represented in the corresponding reference transcriptome. These transcripts were identified by alignment to the reference transcriptomes using BLAT (version 2.17.0) (Kent, 2002) with sensitive parameters (-tileSize=8 -maxIntron=0). The maximum intron length was set to zero to enforce contiguous transcript-to-transcript alignments and prevent spliced mappings. Only alignments with an F-score greater than 0.5 were retained.

The performance of CATS-rf, RSEM-EVAL, and TransRate was quantified with the Spearman correlation coefficient between transcript quality scores and F-scores. Transcript-level performance was evaluated by computing the stated correlations within each transcriptome assembly. Furthermore, the association between mean transcript quality scores and mean transcript F-scores was examined across assemblies. This analysis assessed the ability of transcript quality metrics produced by each tool to rank datasets based on the absolute accuracy of the assembled transcripts across heterogeneous libraries. Assembly-level correlations were compared using the two-tailed Meng's z-test for overlapping correlations (R package cocor, version 1.1.4) (Diedenhofen & Much, 2015), with confidence intervals for the correlation coefficient calculated using Fisher's z-transformation (R package DescTools, version 0.99.60) (Signorell, 2025). Additionally, the ability of CATS-rf to capture common assembly errors was evaluated by comparing CATS-rf assembly scores and transcript-level score components across assemblies with increasing levels of introduced mutations.

To evaluate the effect of parameter variation on CATS-rf score components, a random grid search analysis was conducted using 16 transcriptomes assembled from the first replicate of realistic simulations. The dataset contained rnaSPAdes assemblies from four species (*S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *A. thaliana*), each simulated at four mean coverage levels (20×, 40×, 60×, and 80×). Assemblies were evaluated using CATS-rf with default and 20 randomly sampled parameter combinations defined by the grid search. Parameter settings were varied using the following ranges and step sizes: rolling window length for LCR detection (5–30, step 1), mean rolling window coverage threshold for LCR definition (1.5–5, step 0.5), LCR extension penalty (0.1–1, step 0.05), base coverage weight (1–5, step 0.5), rolling window length for LAR detection (5–30, step 1), mean rolling window accuracy threshold (0.8–1, step 0.01), LAR extension penalty (0.005–0.5, step 0.005), lower multiplicative factor (2–12, step 0.5), higher multiplicative factor (4–16, step 0.5), correction factor (1–10, step 1), compression factor 1 (2–10, step 0.5), and compression factor 2 (0.5–7, step 0.5). Correlation between CATS-rf score components and transcript F-scores were computed within each assembly.

All CATS-rf runs were performed using CATS-rf version 1.0.0 with default parameters. The performance of CATS-rf was analysed using R (version 4.4.2) with the following packages used for data processing and visualization: data.table (version 1.16.4) (Barrett et al., 2025), ggplot2 (version 3.5.1), ggcorrplot (version 0.1.4.1) (Kassambara, 2023), cowplot (version 1.1.3) (Wilke, 2025), and ggbiplot (version 0.6.2) (Vu & Friendly, 2024).

3.2.5. CATS-rb Benchmarking

The performance of CATS-rb was evaluated using the previously described set of simulated transcriptome assemblies and assemblies generated from publicly available RNA-seq libraries. Reference genomes and genomic annotations utilized in CATS-rb benchmark were obtained from Ensembl (Table 7). Run parameters were defined according to species-specific characteristics (Table 8). CATS-rb was first applied to the complete assembly set of each species from (i) controlled and (ii) realistic simulations, each paired with reference transcriptomes. Relative transcript and exon scores were analysed for their association with library characteristics.

Table 7. Reference genomes and the corresponding genomic annotation used in CATS-rb benchmarking. Reproduced from Bodulić & Vlahoviček, 2026.

Species	Build	Ensembl release	N scaffolds	Total length (Mb)
<i>S. cerevisiae</i>	R64-1-1	111	17	12.2
<i>C. elegans</i>	WBcel235	111	7	100.3
<i>D. melanogaster</i>	BDGP6.46	111	1870	143.7
<i>A. thaliana</i>	TAIR10	56	7	119.7
<i>M. musculus</i>	GRCm39	111	61	2728.2
<i>H. sapiens</i>	GRCh38	111	194	3099.8

Mb = megabase

Table 8. Species-specific CATS-rb parameters used in benchmarking. Adapted from Bodulić & Vlahoviček, 2026.

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>D. melanogaster</i>	<i>A. thaliana</i>	<i>M. musculus</i>	<i>H. sapiens</i>
Max gene length (bp)	25000	100000	500000	50000	3000000	3000000
Spaln species preset	saccere	caeneleg	drosmela	arabthal	mus_musc	homosapi
Min exon identity	0.98	0.98	0.98	0.98	0.998	0.998
Max intron length (bp)	1500	20000	90000	5000	600000	400000
Min exon set length (bp)	0	0	0	0	125	100
Min transcript set length (bp)	100	150	200	100	450	300
Max transcript set length (bp)	25000	100000	500000	50000	3000000	3000000

bp = base pairs

The stability of relative transcript and exon scores was further analysed by running CATS-rb on five transcriptome assembly subsets. These subsets excluded reference transcriptomes and consisted of assemblies from controlled simulations with progressively lower transcript coverage. Each subset was defined by the coverage level of its highest-coverage assemblies: 51–100×, 31–40×, 11–20×, 5–10×, and 1–4×. Robustness of relative scores was assessed by examining their correlation with annotation-based scores. The stated correlations were also evaluated using realistically simulated and public datasets, where CATS-rb was run separately for each library alongside the corresponding reference transcriptome.

Furthermore, the sensitivity of CATS-rb completeness metrics was directly tested on transcriptome assemblies subjected to multiplicative mutations. Each evaluated set included the native assembly and triplicates of assemblies with six increasing levels of mutation intensity.

Finally, to evaluate the ability of CATS-rb to detect chimerism as a form of structural inconsistency, chimeric transcripts were simulated by randomly merging pairs of transcripts from the reference transcriptomes of the six analysed taxa. Before merging, random deletions were introduced at the fusion points in both transcripts, removing a randomly selected proportion ranging from 0% to 3% of each transcript's length. The resulting assemblies contained 10% chimeric transcripts.

All CATS-rb runs were performed using CATS-rb version 1.0.0 with default parameters. The performance of CATS-rb was evaluated using R (version 4.4.2) with the following packages used for data processing and visualization: data.table (version 1.16.4), ggplot2 (version 3.5.1), and cowplot (version 1.1.3).

3.3. *Sponge RNA-seq Transcriptome Analysis*

Following its development and benchmarking, the CATS framework was applied to RNA-seq data from sponge species, combining reference-free and reference-based assembly evaluation to identify high-quality transcriptome assemblies for downstream lncRNA identification. This section describes the retrieval, preprocessing, and assembly of sponge RNA-seq datasets, along with quality control of the resulting assemblies. This formed the transcriptomic foundation for subsequent lncRNA identification and comparative analyses.

3.3.1. *Sponge RNA-seq Library Retrieval*

A total of 259 RNA-seq libraries from 14 sponge species were retrieved from SRA with the SRA Toolkit (version 3.2.0), spanning 10 Demospongiae species, two Homoscleromorpha species, and single representatives of Hexactinellida and Calcarea (Table 9). Details of the included libraries are presented in the Appendix. Library selection criteria required the availability of a contiguous reference genome for each species, as well as at least one stranded paired-end RNA-seq library with ≥30% of processed reads successfully mapping to the reference genome.

These criteria ensured reliable data for the identification and annotation of lncRNA transcripts.

Table 9. Overview of sponge species selected for lncRNA identification and annotation.

Species	Class	Accession number	N genomic scaffolds	N ₅₀ (Mb)	N RNA libraries
<i>Amphimedon queenslandica</i>	Demospongiae	GCA_049114765.1	3871	0.95	7
<i>Aphrocallistes beatrix</i>	Hexactinellida	GCA_947044365.1	28	5.0	4
<i>Aplysina aerophoba</i>	Demospongiae	GCA_964213935.1	158	6.6	88
<i>Corticium candelabrum</i>	Homoscleromorpha	GCA_963170055.1	115	8.5	2
<i>Crambe crambe</i>	Demospongiae	GCA_949361645.1	124	7.7	2
<i>Dysidea avara</i>	Demospongiae	GCA_964030475.1	108	40.5	37
<i>Ephydatia muelleri</i>	Demospongiae	GCA_949841015.1	165	8.90	13
<i>Halisarca caerulea</i>	Demospongiae	GCF_964019385.1	488	2.8	25
<i>Oscarella lobularis</i>	Homoscleromorpha	GCA_049114765.1	21	3.5	1
<i>Petrosia ficiformis</i>	Demospongiae	GCA_947507565.1	35	9.90	1
<i>Spongia officinalis</i>	Demospongiae	GCA_963678975.2	105	28.5	21
<i>Spongilla lacustris</i>	Demospongiae	GCA_963422355.1	70	10.1	3
<i>Sycon ciliatum</i>	Calcarea	GCA_963281255.1	618	39.0	42
<i>Tethya wilhelma</i>	Demospongiae	GCA_963924555.1	557	6.7	13

Mb = megabase

3.3.2. Sponge Transcriptome Assembly and Evaluation

All RNA-seq libraries were subjected to quality control and preprocessing prior to downstream analyses. Stranded libraries were first standardized to RF orientation. Adapter sequences were removed and low-quality bases were trimmed using Trim Galore (version 0.6.1) alongside cutadapt (version 4.6) with the parameters --stringency 5 -q 5. Subsequently, error correction was performed using Rcorrector (version 1.0.7), which corrects likely sequencing errors by leveraging k-mer frequency distributions. Specifically, low-abundance k-mers are corrected based on their similarity to high-frequency counterparts, improving overall read accuracy (Song & Florea, 2015). Following correction, reads were aggregated and subjected to in silico normalization by Trinity (version 2.15.1) with a maximum coverage threshold of 50× and strand-aware parameters (--max_cov 50 --SS_lib_type RF). The normalization step aims to reduce redundancy by downsampling reads from highly expressed transcripts while preserving transcriptome diversity (Grabherr et al., 2011).

Normalized reads were assembled into transcripts using seven de novo assembly strategies for each species: rnaSPAdes (version 3.15.4) with default parameters, rnaSPAdes with five alternative k-mer configurations, and Trinity (version 2.15.1) with default parameters. The alternative rnaSPAdes k-mer sets were (i) 21, 31, 41, 51; (ii) 21, 31, 41, 51, 61, 71; (iii) 21, 41, 61, 81; (iv) 31, 41, 51, 61, 71; and (v) 31, 51, 71, 91. All assemblies were performed in strand-aware mode.

De novo transcriptome assembly quality was evaluated using CATS-rf (version 1.0.4) with strand-aware settings (-S rf), while relative completeness was assessed using CATS-rb (version

1.0.3) against the respective reference genome. CATS-rb was run using the complete set of de novo assemblies for each species with strand-aware mapping, *A. queenslandica* mapping preset, a maximum intron length of 10 kb (kilobase), and a maximum gene size of 500 kb (-S rf -p amphquee -i 10000 -m 500000). Assemblies were compared using six complementary performance metrics: N₅₀, proportion of RNA-seq reads successfully mapping to the assembly, CATS-rf assembly score, mean fraction of transcript aligned to the genome, CATS-rb relative transcript score, and CATS-rb relative exon score. These metrics captured two major aspects of assembly performance: transcript quality (N₅₀, genome alignment fraction, and CATS-rf score) and assembly completeness (read mapping proportion, relative transcript score, and relative exon score). For each species, z-scores were calculated for every metric across all assemblies. Mean z-scores were then computed separately for transcript quality and completeness metrics, and overall assembly quality was determined as the average of these two category scores. Assembly quality metrics were analysed in R (version 4.5.1) with packages data.table (version 1.17.8), ggplot2 (version 4.0.2), and pheatmap (version 1.0.13) (Kolde, 2025).

Normalized reads were also assembled using Trinity's genome-guided approach. Reads were first aligned to the reference genome using HISAT2 (version 2.2.1) (Kim et al., 2019) with strand-aware parameters, a maximum intron length of 10 kb, and settings optimized for downstream transcriptome assembly (--rna-strandness RF --max-intronlen 10000 --dta). These alignments were used for genome-guided Trinity assembly, also employing strand-aware settings and a maximum intron size of 10 kb (--SS_lib_type RF --genome_guided_max_intron 10000).

To further improve transcriptome assembly quality, the best-performing rnaSPAdes assembly (as determined by the mean z-score), the Trinity de novo assembly, and the genome-guided Trinity assembly were merged within each species. Redundancy within merged assemblies was reduced using CD-HIT-EST (version 4.8.1). This tool performs redundancy reduction by identifying similar nucleotide sequences using short-word filtering followed by pairwise sequence comparison, clustering sequences above a specified identity threshold and retaining a single representative for each cluster (Fu et al., 2012). Four clustering stringencies were used: (i) -c 0.98 -n 10, (ii) -c 0.96 -n 10, (iii) -c 0.93 -n 9, and (iv) -c 0.90 -n 8, where -c specifies the minimum sequence identity required for clustering and -n defines the word size used during similarity searches. Clustering was performed in strand-aware mode (-r 0) with sequences sorted by descending length prior to clustering. The resulting non-redundant assemblies were evaluated alongside the original de novo and genome-guided assemblies using CATS-rb with parameters analogous to those used for the de novo assembly sets (-S rf -p amphquee -i 10000 -m 500000). The best assembly for each species was selected based on the CATS-rb relative transcript score. Among clustered assemblies, less stringent clustering levels were favored when their relative transcript scores were not substantially lower than those of more stringent alternatives, applying a threshold difference of 0.03. Similarly, the best-performing unclustered assembly was preferred when its CATS-rb transcript score was

within 0.03 of the highest-scoring clustered assembly. The selected assembly was designated as the final optimized transcriptome and was subsequently used for lncRNA identification and annotation. CATS-rb transcript scores were visualized in R (version 4.5.1) using the pheatmap package (version 1.0.13).

3.4. *Genome Decontamination*

The transcriptome assemblies generated in the preceding stage provided the foundation for subsequent lncRNA characterization. However, the accuracy of these analyses also depended on the quality of the corresponding reference genomes. Sponge genome assemblies frequently contain contaminating sequences originating from associated microorganisms (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021), which can interfere with transcript alignment, gene annotation, and lncRNA prediction. Therefore, genome decontamination was performed prior to lncRNA identification to ensure that downstream analyses were restricted to sponge-derived sequences. This chapter is adapted from Bodulić & Vlahoviček (2026), in which the genome decontamination pipeline was originally described.

3.4.1. *Genome Decontamination Pipeline*

The genome decontamination pipeline developed within this dissertation combines multiple lines of evidence to provide an integrated contamination decision and assign taxonomic annotations to contaminant scaffolds (Figure 10). The pipeline takes as input a genome assembly together with the target taxon label. By default, only scaffolds longer than 1 kb are analysed.

The decontamination pipeline calculates three binary scores for each genomic scaffold: a compositional score computed from k-mer frequencies and GC content, a protein score based on taxonomic assignments of encoded proteins, and a nucleotide score derived from taxonomic k-mer classification. Contaminant scaffolds are identified based on the combined outcomes of these scoring components. Detailed descriptions of each scoring metric are provided below.

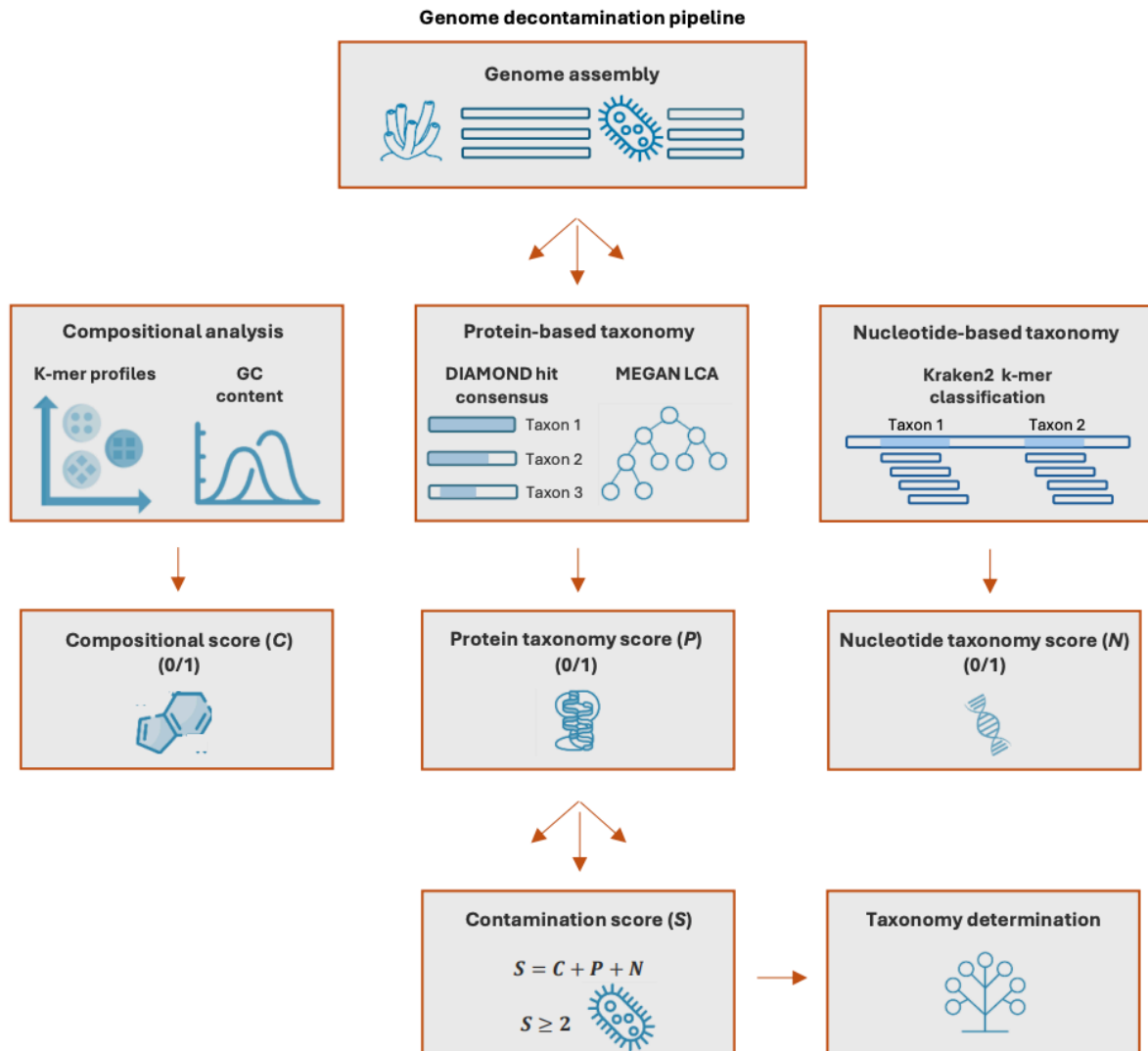


Figure 10. Overview of the genome decontamination pipeline. The pipeline evaluates each genomic scaffold using three independent sources of evidence. This includes (i) compositional outlier identification using k-mer frequencies and GC content (compositional score), (ii) protein-level taxonomic classification using the DIAMOND hit consensus method and MEGAN (protein taxonomy score), and (iii) nucleotide-level taxonomic classification using Kraken2 (nucleotide taxonomy score). Binary scores are combined into a final contamination score, with scaffolds classified as contaminants when at least two lines of evidence imply contamination. Contaminant scaffolds are assigned taxonomic labels based on the available nucleotide- or protein-based evidence. LCA = lowest common ancestor, C = compositional score, P = protein taxonomy score, N = nucleotide taxonomy score, S = contamination score. Reproduced from Bodulić & Vlahoviček, 2026.

The compositional score detects genomic scaffolds exhibiting atypical k-mer profiles or GC content. Scaffold sequences are first partitioned into sliding k-mers of length 5, with each k-mer collapsed with its reverse complement into a single canonical form defined as the lexicographically smaller sequence (R package Biostrings). For each scaffold, k-mer counts are converted to proportions and transformed using the centered log-ratio to correct for compositional effects inherent to relative frequency data (R package compositions) (van den Boogaart et al., 2025). Principal component analysis (PCA) is performed on the transformed k-mer frequencies, retaining the top principal components that together explain at least 90% of the frequency variance, up to a maximum of 20 components.

The resulting principal components are used as input for unsupervised clustering via Gaussian finite mixture models (R package mclust) (Scrucca et al., 2023). Models are fitted across a range of cluster numbers (from $N = 1$ up to a predefined maximum) and evaluated under six covariance structures (EII, EEI, EEE, VII, VEI, and VVI). The optimal model is selected based on the integrated completed likelihood criterion. Following model selection, the reference cluster is defined as the cluster with the greatest cumulative scaffold length, under the assumption that these scaffolds correspond to the target genome. Mahalanobis distances are then computed between each cluster centroid and the centroid of the reference cluster. Clusters are classified as compositional outliers if the square root of this distance exceeds a threshold derived from the chi-square distribution, with degrees of freedom equal to the number of retained principal components. All scaffolds assigned to such clusters are subsequently flagged as compositional outliers.

Following k-mer decomposition, GC content is calculated for each scaffold as the proportion of guanine and cytosine bases within the sequence. Deviations from the genome-wide mean are quantified using z-scores. Scaffolds with absolute z-scores exceeding a predefined threshold are classified as GC outliers. The compositional score (C) is defined as a binary indicator, taking a value of 1 if a scaffold is identified as an outlier based on either k-mer profiles or GC content, and 0 if no compositional deviation is detected.

Protein-based identification of taxonomic outliers is conducted by integrating two complementary strategies: a custom DIAMOND consensus method and a lowest common ancestor (LCA) approach implemented in MEGAN. Each genomic scaffold is first queried against the National Center for Biotechnology Information (NCBI) non-redundant protein database (Sayers et al., 2026) using DIAMOND BLASTx. DIAMOND performs translated sequence similarity searches by aligning six-frame translations of nucleotide sequences against protein databases (Buchfink et al., 2015). The employed parameters (`--sensitive -e 1e-5 --range-culling -F 15 --top 10`) provide increased alignment sensitivity, apply an E-value threshold of 10^{-5} , reduce redundant alignments through range culling, allow frameshift-aware translated alignments with a frameshift penalty of 15, and retain only hits scoring within 10% of the best match. Resulting subject hits are assigned taxonomic identifiers (R package taxonomizr) (Sherrill-Mix, 2025) and are subsequently grouped into one of 14 kingdom-level

categories (Bacteria, Archaea, Viridiplantae, Haptophyta, Rhodophyta, Stramenopiles, Alveolata, Excavata, Rhizaria, Choanoflagellata, Amoebozoa, Fungi, Metazoa, and viruses). Overlapping DIAMOND hits along genomic coordinates are merged into taxon-specific DIAMOND hit sets (R package GenomicRanges). Each resulting hit set is assigned a weight S_w calculated as:

$$S_w = \frac{\sum_1^{N_h} B}{N_h} * \sqrt{S_l}$$

Here, N_h is the number of DIAMOND hits per hit set, B denotes the hit bitscore and S_l represents the hit set length. This formulation ensures that hit sets are weighted according to alignment quality and length.

Each scaffold-taxon combination is assigned a weight W given by:

$$W = \sum_1^{N_{s,t}} S_w$$

where $N_{s,t}$ denotes the number of hit sets assigned to taxon t on scaffold s .

Within the DIAMOND consensus framework, a scaffold is classified as a taxonomic outlier when the maximum weight of any non-target taxon exceeds the target taxon weight multiplied by a specified factor. Additionally, the total length of the highest-weight non-target taxon hit sets must exceed a minimum fraction of the scaffold length and must be greater than the total hit set length of the target taxon.

Each scaffold is independently assigned a taxonomic label using MEGAN (Huson et al., 2018) on the DIAMOND search output, with parameters optimized for long sequences and an E-value threshold of 10^{-5} (-lg -alg longReads -me 1e-5). MEGAN performs taxonomic classification by analyzing sequence similarity matches and assigning each query to the lowest common ancestor (LCA) of the taxa represented among its significant database hits. Scaffolds classified as belonging to a non-target taxon by the LCA algorithm are designated as outliers. The protein taxonomy score (P) is defined as a binary variable equal to 1 if a scaffold is identified as an outlier by either the DIAMOND consensus or MEGAN, and equal to 0 if no evidence of taxonomic inconsistency is detected.

Nucleotide-based detection of taxonomic outliers is performed using Kraken2 (Wood et al., 2019), with scaffolds queried against the NCBI core nucleotide database (Sayers et al., 2026) using default parameters. Kraken2 determines scaffold taxonomy by classifying sliding k-mers based on exact matches to database k-mers. Classified k-mers are mapped to taxonomic identifiers and aggregated into the same set of 14 kingdom-level taxonomic groups utilized in the DIAMOND consensus method.

For each scaffold–taxon combination, an absolute k-mer fraction A is computed as:

$$A = \frac{N_t}{N_s}$$

where N_t represents the number of k-mers assigned to a taxon and N_s denotes the total number of k-mers per scaffold.

A scaffold is identified as a nucleotide-based taxonomic outlier if the following criteria are met: (i) the maximum absolute k-mer fraction among non-target taxa exceeds the target taxon k-mer fraction multiplied by a specified factor and (ii) the maximum non-target k-mer fraction exceeds a specific absolute threshold. The nucleotide taxonomy score (N) is defined as a binary indicator equal to 1 when a scaffold is classified as a nucleotide-based taxonomic outlier and 0 when no evidence of taxonomic discrepancy is detected.

The scaffold contamination score (S) integrates evidence from compositional, protein, and nucleotide scoring components. For each scaffold, the score is calculated as:

$$S = C + P + N$$

A scaffold is considered to originate from a contaminant when $S \geq 2$, reflecting support from at least two independent lines of evidence.

Scaffolds classified as contaminants are assigned a taxonomic label at the highest possible resolution based on either nucleotide- or protein-based evidence. When nucleotide-based detection indicates a taxonomic signal ($N = 1$), scaffolds are assigned to the non-target taxon exhibiting the highest number of matching k-mers at a given taxonomic rank. Taxonomic assignment is performed hierarchically, beginning at the genus level and proceeding to higher ranks until the following conditions are met: (i) the absolute k-mer fraction for the taxon exceeds a predefined minimum threshold and (ii) relative k-mer fraction for the same taxon surpasses a second threshold. The relative k-mer fraction R is calculated as:

$$R = \frac{N_t}{N_c}$$

where N_c represents the total number of taxonomically classified k-mers per scaffold.

When nucleotide-based evidence is not detected ($N = 0$), or when the above criteria are not satisfied at any taxonomic rank, contaminant taxonomy is determined using protein-based evidence. In this case, the scaffold is assigned to the taxon with the highest cumulative length of associated DIAMOND hit sets. Taxonomic ranks are again assessed hierarchically until the following criteria are fulfilled: (i) the cumulative length of hit sets assigned to a given taxon surpasses a minimum fraction of the scaffold length and (ii) exceeds a minimum fraction of

the total hit set length per scaffold. If the described conditions are not met for any taxonomic rank, the scaffold is assigned an unknown taxonomy. This hierarchical methodology prioritizes nucleotide-based evidence when available, while utilizing protein alignment to provide taxonomic resolution for scaffolds that lack sufficient nucleotide signal.

The decontamination pipeline comprises tunable parameters that were not explicitly optimized for the datasets analysed in this work. Rather, default parameter values were selected to reflect biologically reasonable settings (Table 10)

Table 10. Default values of tunable parameters included in the genome decontamination pipeline. Adapted from Bodulić & Vlahoviček, 2026.

Parameter	Default value	Rationale
Maximum number of Gaussian finite mixture model clusters	10	The default value enables sufficient flexibility to detect compositional heterogeneity while avoiding unnecessary model complexity. Assemblies typically comprise a small number of compositional groups (e.g., target genome, organelles, contaminants), and a maximum of 10 clusters offers adequate capacity to represent this diversity
Chi-square significance level threshold for outlier cluster detection	0.005	The default level was chosen to reduce false positives arising from the expected variation in k-mer composition while preserving sensitivity to stronger deviations indicative of contamination. A more stringent threshold than commonly used significance levels is adopted as no formal correction for multiple comparisons is applied
GC content z-score threshold for outlier scaffold detection	2.5	A conservative default value was selected to minimize erroneous outlier calls due to intrinsic genomic variation
Multiplying factor for the DIAMOND hit set weight comparison	1.2	The default multiplying factor ensures that non-target taxonomic signal exceeds the target signal by a significant margin, reducing sensitivity to minor fluctuations in alignment scores while preserving the ability to detect genuine contamination
Minimum non-target DIAMOND hit set coverage fraction of a scaffold required for contaminant classification	0.2	The default coverage fraction requires that non-target DIAMOND hit sets span a significant fraction of the scaffold, reducing the effect of isolated alignments while retaining sensitivity to contaminant scaffolds with a low number of PCGs
Multiplying factor for the Kraken2 k-mer ratio comparison	1.2	This parameter follows the same rationale as the multiplying factor for the DIAMOND hit set weight comparison
Minimum non-target Kraken2 k-mer fraction required for contaminant classification	0.05	The default value requires that k-mers derived from non-target taxa constitute a significant fraction of the scaffold, while ensuring sensitivity to taxa missing from the NCBI core nucleotide database
Minimum Kraken2 k-mer fraction required to assign a taxon to a scaffold	0.1	These thresholds ensure that taxonomic assignment is supported by absolute k-mer fraction and relative signal compared to other taxa. The minimum k-mer fraction ensures that a significant portion of the scaffold supports the assignment, while the relative fraction requires that the assigned taxon constitutes the dominant signal
Minimum Kraken2 relative k-mer fraction required to assign a taxon to a scaffold	0.5	
Minimum DIAMOND hit set coverage fraction required to assign a taxon to a scaffold	0.25	These parameters follow the same rationale as nucleotide-based taxonomy assignment thresholds
Minimum DIAMOND hit set relative coverage fraction required to assign a taxon to a scaffold	0.5	

PCG = protein-coding gene, NCBI = National Center for Biotechnology Information

3.4.2. Genome Decontamination Pipeline Validation

To assess genome decontamination pipeline performance, a spike-in validation dataset was created utilizing bona fide sponge genomic sequences and a representative collection of contaminant genomes. Sponge sequences were derived from chromosome-level genome assemblies spanning 18 species not included in the lncRNA characterization (Table 11). For each sponge genome, only chromosomal sequences were included, while unlocalized scaffolds were removed based on genome annotation. In addition, the sponge dataset was supplemented by mitochondrial sequences from *Bolosoma cyanae*, *Chondrosia reniformis*, *Petrosia crassa*, *Phakellia ventilabrum*, and *Weberella bursa*.

Additionally, a diverse panel of contaminant taxa was incorporated to reflect typical sponge symbionts and environmental organisms across a broad taxonomic range. This included genome assemblies from 71 species representing bacteria, archaea, protozoa, algae, fungi, and viruses (Table 12). The dataset also contained five plasmid sequences from the following bacteria: *Desulfobacter hydrogenophilus* (RefSeq NZ_CP036314.1), *Paludibaculum fermentans* (RefSeq NZ_CP063850.1), *Pseudovibrio brasiliensis* (RefSeq NZ_CP074127.1), *Synechococcus elongatus* (RefSeq NZ_CP085786.1), and *Thalassospira xiamenensis* (RefSeq NZ_CP004389.1).

Table 11. Sponge genome assemblies included in the spike-in validation dataset.

Species	Class	GenBank accession number	N scaffolds	N ₅₀ (Mb)
<i>Agelas tubulata</i>	Demospongiae	GCA_964245335.1	56	12.1
<i>Aiolochoira crassa</i>	Demospongiae	GCA_964263335.1	37	7.9
<i>Aphrocallistes vastus</i>	Hexactinellida	GCA_963281255.1	28	5.0
<i>Aplysina cauliformis</i>	Demospongiae	GCA_965643675.1	50	8.2
<i>Aplysina cavernicola</i>	Demospongiae	GCA_964659575.1	23	7.6
<i>Bolosoma cyanae</i>	Hexactinellida	GCA_964258925.1	27	4.8
<i>Chondrilla caribensis</i>	Demospongiae	GCA_964020155.1	46	5.3
<i>Chondrosia reniformis</i>	Demospongiae	GCA_947172415.1	14	8.5
<i>Geodia parva</i>	Demospongiae	GCA_964274975.1	48	4.9
<i>Geodia phlegraei</i>	Demospongiae	GCA_964341465.1	42	5.0
<i>Halichondria panicea</i>	Demospongiae	GCA_963675165.1	39	7.1
<i>Oscarella pearsei</i>	Homoscleromorpha	GCA_947507565.1	21	3.5
<i>Pachastrella ovisternata</i>	Demospongiae	GCA_964263345.1	26	7.2
<i>Petrosia crassa</i>	Demospongiae	GCA_965785925.1	35	9.90
<i>Phakellia ventilabrum</i>	Demospongiae	GCA_964276705.1	26	8.4
<i>Spongia lamella</i>	Demospongiae	GCA_963930835.2	24	14.5
<i>Weberella bursa</i>	Demospongiae	GCA_965644515.1	24	3.8
<i>Xestospongia bergquistia</i>	Demospongiae	GCA_963965975.1	20	8.9

Mb = megabase

Table 12. Contaminant genome assemblies included in the spike-in validation dataset.

Species	Class	Genbank accession number	N scaffolds	N ₅₀ (Mb)
<i>Leisingera methylohalidivorans</i>	Alphaproteobacteria	GCA_000511355.1	3	4.1
<i>Phaeobacter gallaeciensis</i>	Alphaproteobacteria	GCA_000511385.1	8	3.8
<i>Pseudovibrio brasiliensis</i>	Alphaproteobacteria	GCA_018282095.1	6	4.8
<i>Pseudovibrio denitrificans</i>	Alphaproteobacteria	GCA_047444875.1	3	4.9
<i>Roseobacter litoralis</i>	Alphaproteobacteria	GCA_000154785.2	4	4.5
<i>Ruegeria pomeroyi</i>	Alphaproteobacteria	GCA_000011965.2	2	4.1
<i>Sulfitobacter pontiacus</i>	Alphaproteobacteria	GCA_040790665.1	5	3.0
<i>Sulfitobacter rhodophyticola</i>	Alphaproteobacteria	GCA_041281685.2	6	4.0
<i>Thalassospira xiamenensis</i>	Alphaproteobacteria	GCA_000300235.2	2	4.6
<i>Alteromonas naphthalenivorans</i>	Gammaproteobacteria	GCA_000213655.1	1	5.0
<i>Alteromonas stellipolaris</i>	Gammaproteobacteria	GCA_001562195.1	1	4.6
<i>Marinobacter nauticus</i>	Gammaproteobacteria	GCA_000284615.1	1	4.0
<i>Pseudomonas putida</i>	Gammaproteobacteria	GCA_000412675.1	1	6.2
<i>Vibrio vulnificus</i>	Gammaproteobacteria	GCA_002224265.1	2	3.3
<i>Desulfobacter hydrogenophilus</i>	Deltaproteobacteria	GCA_004319545.1	3	5.2
<i>Desulfobacter postgatei</i>	Deltaproteobacteria	GCA_000233695.3	1	4.0
<i>Desulfobulbus propionicus</i>	Deltaproteobacteria	GCA_000186885.1	1	3.9
<i>Desulfovibrio mangrovi</i>	Deltaproteobacteria	GCA_026230175.1	1	3.9
<i>Geobacter sulfurreducens</i>	Deltaproteobacteria	GCA_000007985.2	1	3.8
<i>Bacillus cabrialesii</i>	Bacilli	GCA_004124315.2	1	4.1
<i>Bacillus velezensis</i>	Bacilli	GCA_000015785.2	1	3.9
<i>Oceanobacillus iheyensis</i>	Bacilli	GCA_000011245.1	1	3.6
<i>Virgibacillus natechei</i>	Bacilli	GCA_026013645.1	1	3.9
<i>Virgibacillus pantothenicus</i>	Bacilli	GCA_018075365.1	1	4.8
<i>Occallatibacter riparius</i>	Acidobacteriia	GCA_025264625.1	1	6.8
<i>Micrococcus luteus</i>	Actinobacteria	GCA_900475555.1	1	2.5
<i>Micrococcus porci</i>	Actinobacteria	GCA_020097155.1	1	2.6
<i>Streptomyces collinus</i>	Actinobacteria	GCA_031348265.1	1	8.9
<i>Leptolyngbya boryana</i>	Cyanophyceae	GCA_002142475.1	4	6.2
<i>Prochlorococcus marinus</i>	Cyanophyceae	GCA_000015665.1	1	1.7
<i>Synechococcus elongatus</i>	Cyanophyceae	GCA_022984195.1	3	2.7
<i>Niabella soli</i>	Chitinophagia	GCA_000243115.3	1	4.7
<i>Echinicola vietnamensis</i>	Cytophagia	GCA_000325705.1	1	5.6
<i>Dehalococcoides mccartyi</i>	Dehalococcoidia	GCA_000011905.1	1	1.5
<i>Dehalogenimonas formicexedens</i>	Dehalococcoidia	GCA_001953175.1	1	2.1
<i>Formosa sediminum</i>	Flavobacteria	GCA_007197735.1	1	3.9
<i>Halobacterium salinarum</i>	Halobacteria	GCA_004799605.1	3	2.2
<i>Halorubrum ruber</i>	Halobacteria	GCA_018228765.1	1	3.0
<i>Nitrospira japonica</i>	Nitrospira	GCA_900169565.1	1	4.1
<i>Nitrospira moscoviensis</i>	Nitrospira	GCA_001273775.1	1	4.6
<i>Haloferula helveola</i>	Planctomycetia	GCA_037076345.1	1	5.7
<i>Planctomyces</i> sp.	Planctomycetia	GCA_001610835.1	1	8.4
<i>Rhodopirellula baltica</i>	Planctomycetia	GCA_000196115.1	1	7.1
<i>Paludibaculum fermentans</i>	Sphingobacteriia	GCA_015277775.1	2	9.5
<i>Verrucomicrobium spinosum</i>	Verrucomicrobiae	GCA_000172155.1	1	8.2
<i>Methanobrevibacter smithii</i>	Methanobacteria	GCA_000016525.1	1	1.9
<i>Nitrosopelagicus brevis</i>	Nitrososphaeria	GCA_000812185.1	1	1.2
<i>Nitrosotenuis cloacae</i>	Nitrososphaeria	GCA_000955905.3	1	1.6
<i>Nitrosopumilus maritimus</i>	Nitrososphaeria	GCA_000018465.1	1	1.6
<i>Cryptosporidium parvum</i>	Conoidasida	GCA_000165345.1	8	1.1

<i>Dictyostelium discoideum</i>	Dictyostelia	GCA_000004695.1	50	5.5
<i>Acanthamoeba castellanii</i>	Discosea	GCA_000313135.1	384	0.3
<i>Capsaspora owczarzaki</i>	Filasterea	GCA_000151315.2	84	1.6
<i>Paramecium tetraurelia</i>	Oligohymenophorea	GCA_000165425.1	697	0.4
<i>Tetrahymena thermophila</i>	Oligohymenophorea	GCA_000189635.1	1150	0.5
<i>Phaeodactylum tricornutum</i>	Bacillariophyceae	GCA_000150955.2	88	1.0
<i>Thalassiosira pseudonana</i>	Bacillariophyceae	GCA_000149405.2	64	2.0
<i>Ostreococcus tauri</i>	Mamiellophyceae	GCA_000214015.2	20	0.8
<i>Chlorella sorokiniana</i>	Trebouxiophyceae	GCA_025917655.1	13	3.0
<i>Aspergillus sydowii</i>	Eurotiomycetes	GCA_001890705.1	97	2.3
<i>Malassezia globosa</i>	Malasseziomycetes	GCA_000181695.2	62	0.7
<i>Candida parapsilosis</i>	Saccharomycetes	GCA_000182765.2	8	2.1
<i>Debaryomyces hansenii</i>	Saccharomycetes	GCA_000006445.2	7	2.0
<i>Spizellomyces punctatus</i>	Spizellomycetes	GCA_000182565.2	38	1.5
Alteromonas phage	Caudoviricetes	GCA_010706295.1	1	0.09
Ruegeria phage	Caudoviricetes	GCA_000925815.1	1	0.06
Vibrio phage VP16T	Caudoviricetes	GCA_004216455.1	1	0.05
Vibrio phage CTXphi	Faserviricetes	GCA_000893195.1	1	0.01
Chaetoceros socialis RNA virus 01	Pisoniviricetes	GCA_000882095.1	1	0.01
Heterosigma akashiwo RNA virus	Pisoniviricetes	GCA_000863545.1	1	0.01
Acanthamoeba polyphaga mimivirus	Megaviricetes	GCA_000888735.1	1	1.2

Mb = megabase

The validation dataset was generated by fragmenting the input sequences to replicate scaffold length distributions observed in the sponge assemblies analysed for contamination. Only sequences longer than 1 kb were considered, as shorter sequences were not classified by the decontamination pipeline. The length distribution of sponge scaffolds analysed for contamination was characterized using 13 predefined length bins: 1–2 kb, 2–4 kb, 4–10 kb, 10–20 kb, 20–30 kb, 30–50 kb, 50–100 kb, 100–200 kb, 200–500 kb, 500 kb–1 Mb (megabase), 1–5 Mb, 5–10 Mb, and 10–100 Mb. The fraction of scaffolds within each bin guided fragment sampling to match empirical assembly profiles. Fragments were produced by sampling contiguous segments from source sequences, with fragment lengths randomly drawn from the length bins. For sponge-derived sequences, unsampled input sequence segments longer than 1 kb were returned to the sampling pool and could be reused up to four times. Fragments containing more than 10% ambiguous bases were excluded from the final set.

The final validation dataset contained 13834 sequences, of which 4653 (33.6%) were derived from contaminants. The N_{50} value of sponge-derived sequences was 7.01 Mb, while contaminant-derived sequences had an N_{50} of 1.68 Mb (Figure 11).

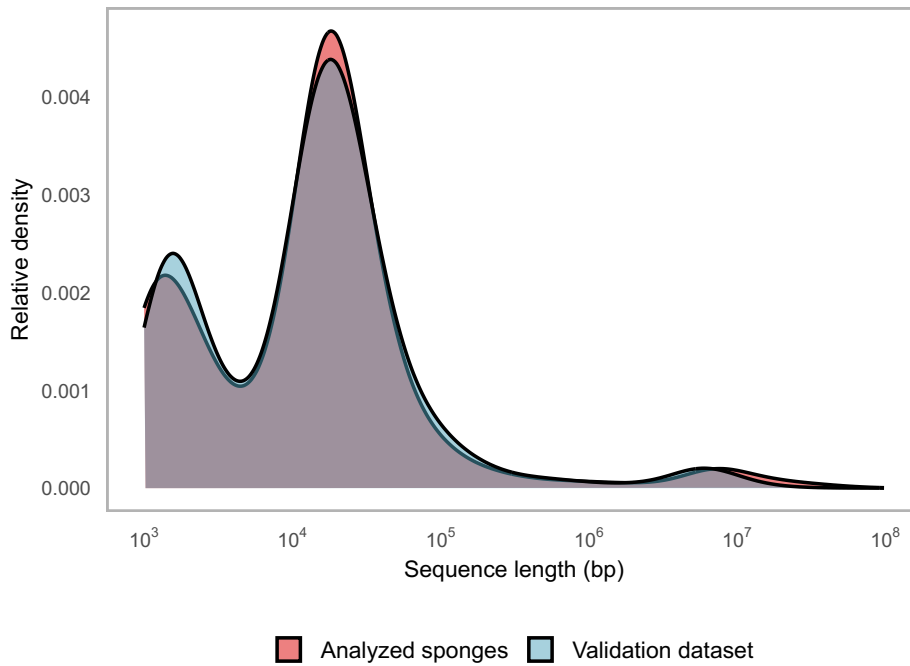


Figure 11. Distribution of sequence length in the spike-in validation dataset and sponge genome assemblies analysed for contamination. Reproduced from Bodulić & Vlahoviček, 2026.

The decontamination pipeline was applied to the validation dataset using default parameters, apart from the maximum number of compositional scaffold clusters. This parameter was increased to 20 to account for the taxonomic species diversity present in the validation dataset, which exceeds that typically expected in a single genome. The following software and database versions were used: DIAMOND (version 2.1.13), NCBI non-redundant database (downloaded on September 17, 2025), MEGAN (version 7.1.1), MEGAN taxonomy mapping file (version 7.0.0), Kraken2 (version 2.1.6), NCBI core nucleotide database (downloaded on October 29, 2025), taxonomizr (version 0.11.1), taxonomizr database (downloaded on December 27, 2025), Biostrings (version 2.76.0), compositions (version 2.0.9), mclust (version 6.1.2), and GenomicRanges (version 1.62.1).

Decontamination results from the validation dataset were analysed to assess the performance of the pipeline and its individual scoring components. Accuracy, precision, and recall were calculated for the compositional, protein, and nucleotide scores, as well as for the final contamination score. The overlap among contamination calls generated by each scoring component was examined using UpSet plots. Contaminant taxa were further compared by hierarchical clustering based on recall values across all scoring components and the integrated score, using complete linkage and Euclidean distance. Finally, the taxonomic classification accuracy of contaminant scaffolds was evaluated by comparing assigned taxonomic labels with the known source taxa of contaminant sequences. Validation dataset generation and result analysis were performed using R (version 4.5.1), with the following R packages used for

data manipulation and visualization: data.table (version 1.17.8), ggplot2 (version 4.0.2), ComplexUpset (version 1.3.3) (Lex et al., 2014), and pheatmap (version 1.0.13).

3.4.3. *Sponge Genome Assembly Decontamination*

The genome decontamination pipeline was applied to 14 sponge genome assemblies utilized for lncRNA characterization. All assemblies were processed using default parameters, with software dependencies and database versions consistent with those used in the validation dataset analysis. Decontamination results were summarized as the proportion of scaffolds and total genome length classified as contaminants. Sponge species and contaminant taxa were compared by hierarchical clustering complete linkage and the Euclidean distance metric. The overlap among contamination calls generated by the scoring components was visualized using Venn diagrams.

To assess potential false-positive contaminant identification, genome assembly completeness was compared between the original and decontaminated genome assemblies. Completeness was assessed using BUSCO (version 6.0.0) (Tegenfeldt et al., 2025) with the Metazoa lineage. Decontamination results were analysed using R (version 4.5.1) with the following packages: data.table (version 1.17.8), ggplot2 (version 4.0.2), pheatmap (version 1.0.13), and ggVennDiagram (version 1.5.4) (Gao & Dusa, 2025).

3.5. *Genome Annotation*

Comprehensive structural and functional annotation of the analysed sponge genomes was essential for downstream lncRNA analyses. The first stage involved repeat characterization, which supported accurate PCG annotation and enabled investigation of the relationships between lncRNA genes and TEs. The second step focused on protein annotation, which facilitated comparisons between lncRNA genes and PCGs, along with the evaluation of potential cis-regulatory roles of lncRNAs.

3.5.1. *Repeat Annotation*

The repeat annotation strategy followed a comprehensive approach designed to generate a high-confidence catalogue of genomic repeats across all analysed sponge genomes. Repeats were first identified de novo in each sponge genome using RepeatModeler (version 2.0.7) with default parameters, including the integrated LTR structural discovery workflow based on LTR_retriever. RepeatModeler performs de novo repeat discovery by combining complementary algorithms to identify and cluster repetitive sequences directly from genomic data based on sequence similarity and repeat structure. These candidate repeats are then refined into consensus family models, classified where possible against known TE databases, and compiled into a custom repeat library for downstream genome masking and annotation (Flynn et al., 2020). RepeatModeler outputs were separated into known and unknown repeat families for each analysed species.

Repeats were also identified de novo in each genome using EDTA (version 2.3.0) with the generic eukaryotic species setting (`--species others`) and the custom repeat library generated by RepeatModeler. The tool performs comprehensive de novo TE annotation by integrating multiple specialised detection tools to identify structurally distinct TE classes, including LTR and non-LTR retrotransposons, DNA transposons, and helitrons, followed by classification and library construction (Ou et al., 2019). EDTA outputs were likewise separated into known and unknown repeat families.

Unknown repeats identified by both RepeatModeler and EDTA were first classified using TEsorter (version 1.5.1) with default parameters and the Metazoan REXdb database of TE protein domains. TEsorter classifies transposable elements by translating repeat sequences and searching for conserved protein domains against the REXdb database using hidden Markov model (HMM)-based searches, enabling phylogeny-informed classification of repetitive elements into established TE superfamilies (Zhang et al., 2022). Remaining unclassified repeats were subsequently analysed using DeepTE (version 10/22/2022), a deep learning-based TE classification tool that utilizes sequence k-mer composition (Yan et al., 2020). A classification probability threshold of 0.75 was applied to prioritise high-confidence assignments.

Known repeat families from all species were merged, length-sorted, and clustered using CD-HIT-EST (version 4.8.1) (parameters `-c 0.9 -n 8`) across six major repeat categories: LTR retrotransposons, non-LTR retrotransposons, DNA transposons, helitrons, satellite repeats, and small RNAs. Remaining unknown repeat families from all sponge species were also merged, length-sorted, and clustered using CD-HIT-EST with the same parameters. This process generated final non-redundant libraries of known and unknown repeat families.

Repeat annotation of each sponge genome was then performed using RepeatMasker (version 4.2.3) (Chen, 2004) through three sequential steps. First, genomes were screened against the Dfam database, which contains profile HMMs of repetitive elements (Hubley et al., 2016). Second, genomes were screened using the previously described known repeat family library and third, using the unknown repeat family library. Together, this strategy produced soft-masked genomes for all 14 analysed sponge species, along with a high-confidence repeat annotation.

For each species, the abundance of individual repeat classes was quantified as both the number of annotated repeat elements and the proportion of decontaminated genome length occupied by each class. Prior to genomic coverage calculations, overlapping repeat annotations belonging to the same repeat class were merged into non-redundant genomic intervals. Repeat annotation statistics were calculated in R (version 4.5.2) using the `data.table` (version 1.17.8) and `GenomicRanges` (version 1.62.1) packages.

3.5.2. Protein-coding Gene Annotation

Protein-coding gene annotation was performed on soft-masked genomes using BRAKER3, an automated genome annotation pipeline that integrates both extrinsic evidence and ab initio gene prediction. BRAKER3 combines RNA-seq alignments and homologous protein evidence to train gene prediction models, using specialized tools to accurately infer exon–intron structures and coding sequences. By leveraging transcriptomic data for expressed gene structures alongside conserved protein homology from related species, BRAKER3 generates high-confidence PCG annotations with improved accuracy (Gabriel et al., 2024).

BRAKER3 (version 3.0.8) was run on each soft-masked genome using default parameters. Transcriptomic evidence was provided in the form of normalized RNA-seq reads aligned to the corresponding genome as part of the genome-guided Trinity workflow (Section 3.3.2). Conserved protein homology evidence was supplied from the Metazoan dataset of OrthoDB (version 12) (Tegenfeldt et al., 2025), enabling BRAKER3 to integrate both transcriptomic and protein-based evidence for the prediction of PCGs. For each species, the numbers of genes, exons, and introns were determined, together with the proportion of decontaminated genome length occupied by each feature. Prior to calculating genomic coverage, overlapping elements of the same type were merged into non-redundant genomic intervals. Annotation statistics were calculated in R (version 4.5.2) using the data.table (version 1.17.8) and GenomicRanges (version 1.62.1) packages.

Functional protein annotation was performed on BRAKER3 output using eggNOG-mapper (version 2.1.13), which assigns orthology-based functional annotations by querying predicted protein sequences against the eggNOG orthologous groups database (Cantalapiedra et al, 2021).

3.6. Long Non-coding RNA Analysis

The preceding analyses generated a comprehensive genomic and transcriptomic resources for each sponge species, including optimized transcriptome assemblies, decontaminated genome assemblies, TE annotations, and PCG annotations. These resources enabled the identification of high-confidence lncRNAs and supported their subsequent analysis.

3.6.1. Long Non-coding RNA Identification

Long non-coding RNA identification was performed using a comprehensive pipeline consisting of several sequential filtering steps designed to retain high-quality lncRNA transcripts (Figure 12). The initial stage leveraged transcript-to-genome mappings generated during the transcriptome quality assessment by CATS-rb. These mappings enabled the removal of unmapped transcripts, which likely corresponded to misassembled sequences or contaminant-derived fragments, as well as transcripts exhibiting structural inconsistencies indicative of low transcript accuracy. Mono-exonic transcripts were subsequently excluded,

as they were potentially representative of fragmented or poorly assembled transcripts arising from de novo assembly. Transcripts shorter than 200 bp were then removed in accordance with the canonical lncRNA length definition. The following step removed transcripts with TPM < 1 across all analysed libraries. Transcript expression was quantified for each library using Kallisto (version 0.50.1) with default parameters and strand-specific settings (--rf-stranded). Such lowly expressed transcripts were presumed to represent transcriptional noise resulting from insufficient read coverage. Together, these procedures constituted the transcript quality filtering phase of the pipeline.

The subsequent steps focused on the removal of transcripts with biological identities inconsistent with lncRNA classification. Known structural RNA families were identified and removed using INFERNA (version 1.1.5), which employs covariance models to query transcripts against the Rfam database (Nawrocki et al., 2009). All transcripts yielding hits with an E-value below 10^{-10} were discarded. Coding potential was then assessed using CPC2 (version 1.0.1), a support vector machine-based classifier that evaluates transcript coding likelihood based on ORF quality, Fickett score, and isoelectric point (Kang et al., 2017). All transcripts classified as coding were removed. Remaining transcripts were further screened for protein homology by querying against the NCBI non-redundant protein database using DIAMOND (version 2.1.13). Transcripts with hits exhibiting an E-value below 10^{-10} and query coverage exceeding 20% were excluded. This was complemented by conserved protein domain searches using HMMER (version 3.3.2), which applies profile HMMs to detect remote homology to known protein families (Eddy, 2011). Transcripts yielding significant hits (E-value < 10^{-10}) were discarded. Finally, transcripts whose exons overlapped protein-coding exons from the BRAKER3 annotation in the sense orientation were removed to exclude any residual mRNAs missed by earlier steps. Transcript filtering within each step was performed using R (version 4.5.1) with data.table (version 1.17.8), Biostrings (version 2.76.0), and GenomicRanges (version 1.62.1) packages.

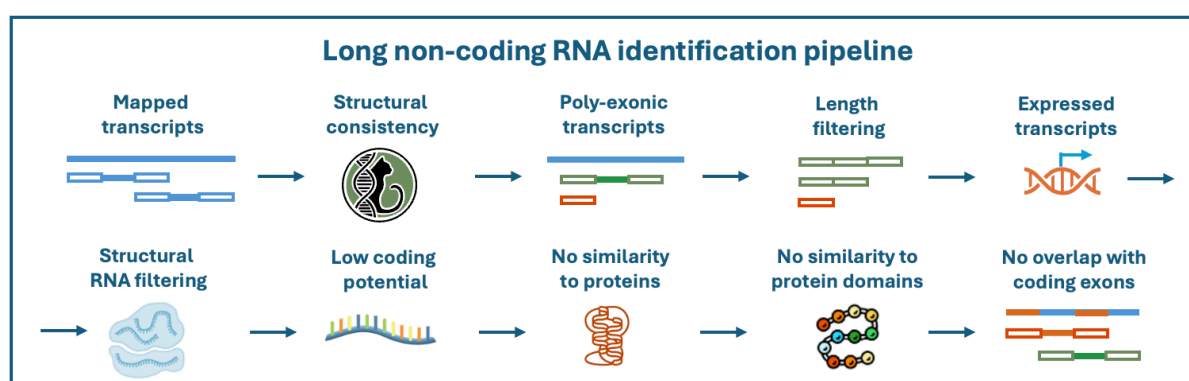


Figure 12. Long non-coding RNA identification pipeline. Overview of the sequential filtering steps used to identify high-confidence lncRNA transcripts from sponge transcriptome assemblies. The pipeline combines transcript quality filtering with removal of known structural and coding RNA transcripts to retain high-fidelity lncRNAs.

3.6.2. General Characterization of Long Non-coding RNAs

Identified lncRNAs were first analysed for their general characteristics. In all analyses, lncRNAs were compared against two distinct transcript groups: PCTs derived from the same transcriptome assemblies and PCTs originating from BRAKER3 genome annotations. Assembly-derived PCTs were defined as the union of transcripts classified as coding by CPC2, transcripts exhibiting sequence similarity to known proteins or conserved protein domains, and transcripts overlapping BRAKER3-defined protein-coding exons in sense.

First, lncRNAs and PCTs were analysed within their genomic context using transcript mappings previously generated with spaln within the CATS-rb pipeline. Exon-based spaln output was aggregated into transcript-level mappings as described in the CATS-rb methodology (Section 3.2.2). For each transcript, only the mapping with the greatest alignment length was retained. Introns were subsequently inferred as genomic gaps between adjacent exons within individual transcript mappings. Gene models were generated separately for lncRNAs and PCTs by strand-aware collapsing of overlapping transcript coordinates into non-redundant genomic loci. Exons and introns shorter than 20 bp were removed from all subsequent analyses.

Lengths of transcripts, genes, exons, and introns were compared among lncRNAs, assembly-derived PCTs, and BRAKER3-derived PCTs. Exon number per transcript and isoform number per gene were similarly compared within each species across the three classes. Isoforms were defined as transcripts overlapping the same genomic locus within each transcript category. All statistical comparisons were performed using the Kruskal–Wallis test followed by Dunn’s post hoc test for pairwise comparisons. P-values were adjusted for multiple testing using the Benjamini–Hochberg method. All tests were two-tailed with statistical significance defined at $p < 0.05$. This analysis was performed using R (version 4.5.1) with the following packages: data.table (version 1.17.8), GenomicRanges (version 1.62.1), Biostrings (version 2.76.0), and ggplot2 (version 4.0.2).

3.6.3. Long Non-coding RNA Splicing Analysis

Donor and acceptor splice-site sequences were extracted from the corresponding genome assemblies based on genomic coordinates. Only splice sites validated by read evidence were included in the analysis. This validation was performed by mapping normalized RNA-seq libraries to the corresponding decontaminated genome assemblies using HISAT2 (version 2.2.1) with strand-specific alignment, a maximum intron length of 10 kb, and non-canonical splice sites scored without penalty (--rna-strandness RF --max-intronlen 10000 --pen-noncansplice 0). Splice junctions were subsequently extracted using regtools (version 1.0.0) (Cotto et al., 2023) with strand-aware processing, requiring a minimum anchor length of 8 bp, and intron lengths between 20 bp and 10 kb (-s RF -a 8 -m 20 -M 10000). Only splice sites showing exact coordinate matches with extracted splice junctions were retained for downstream analyses. Splice-site composition was compared among lncRNAs, assembly-derived PCTs, and BRAKER-derived PCTs. Statistical significance was assessed using chi-square

test, with multiple-testing correction performed using the Benjamini–Hochberg method. All tests were two-tailed with statistical significance defined at $p < 0.05$.

Identified splice junctions were further evaluated based on splice motif strength using MaxEntScan (version 1.0.0). MaxEntScan evaluates splice-site strength using maximum entropy models trained on experimentally validated splice junctions, quantifying how closely a given splice-site sequence conforms to known splicing signals (Yeo & Burge, 2004). The tool was run with default parameters on splice sites supported by RNA-seq reads. For each intron, MaxEntScan scores were averaged across donor and acceptor splice sites. Mean splice site strengths were subsequently compared among lncRNAs, assembly-derived PCTs and BRAKER3-derived PCTs across all analysed species. Statistical significance was assessed using the Kruskal–Wallis test followed by Dunn’s post hoc test, with multiple-testing correction performed using the Benjamini–Hochberg procedure. All statistical tests were two-tailed, with significance defined at $p < 0.05$.

To further characterise splice site composition, sequence logo analysis was performed on canonical GT-AG donor and acceptor splice motifs derived from lncRNAs, assembly-derived PCTs, and BRAKER3-derived PCTs. For each confirmed intron, genomic sequence windows surrounding splice junctions were extracted from the corresponding reference genome while accounting for transcript strand orientation. Donor splice sites were represented by 23 bp windows spanning positions -3 to $+19$ relative to the exon–intron boundary, whereas acceptor splice sites were represented by 23 bp windows spanning positions -19 to $+3$. Nucleotide frequencies at each position were calculated and visualised as sequence logos, with splice motifs aggregated across all analysed sponge species.

Splice site analysis was performed using R (version 4.5.1) with the following packages: data.table (version 1.17.8), GenomicRanges (version 1.62.1), Biostrings (version 2.76.0), ggplot2 (version 4.0.2), and seqLogo (version 1.78.0) (Bembom & Ivanek, 2026).

3.6.4. Long Non-coding RNA Sequence Composition Analysis

To compare sequence composition among lncRNAs, assembly-derived PCTs and BRAKER3-derived PCTs, k-mer-based transcript clustering was performed following the same principles used for calculation of the compositional score in the genome decontamination pipeline (Section 3.4.1). Transcript sequences longer than 400 bp were decomposed into overlapping k-mers ($k = 5$). For each transcript, k-mer counts were converted to relative frequencies and transformed using the centered log-ratio transformation. Following transformation, PCA was performed on the transformed k-mer frequencies, retaining the top principal components explaining at least 90% of the total variance, up to a maximum of 20 components. The resulting principal components were subsequently visualised using Uniform Manifold Approximation and Projection (UMAP) to assess clustering patterns among transcript classes.

GC content was calculated for each transcript as the proportion of guanine and cytosine nucleotides within the sequence. GC content distributions were compared among lncRNAs,

assembly-derived PCTs, and BRAKER3-derived PCTs in each species. Statistical significance was assessed using the Kruskal–Wallis test followed by Dunn’s post hoc test, with multiple-testing correction performed using the Benjamini–Hochberg procedure. All tests were two-tailed, with statistical significance defined at $p < 0.05$. Sequence composition analysis was performed in R (version 4.5.1) with the following packages: data.table (version 1.17.8), Biostrings (version 2.76.0), compositions (version 2.0.9), umap (version 0.2.10) (Konopka, 2023), and ggplot2 (version 4.0.2).

3.6.5. Classification of Long Non-coding RNAs Relative to Protein-coding Genes

Identified lncRNA transcripts were categorised based on their genomic relationship to BRAKER3-annotated PCGs. Following the classification framework of Liu et al. (2021) (Section 2.4.3), lncRNAs were assigned to five categories: intergenic, intronic, antisense, bidirectional, and overlapping. Sense lncRNAs were excluded from all analyses, as these transcripts could not be reliably distinguished from residual PCTs and were filtered by the lncRNA identification pipeline. A distance threshold of 2 kb was used for classification of bidirectional lncRNAs. Transcript, gene, exon, and intron lengths were compared among lncRNA classes using the Kruskal–Wallis test followed by Dunn’s post-hoc test. Isoform number distributions in the analysed lncRNA classes were compared using the same approach. Multiple-testing correction was performed using the Benjamini–Hochberg procedure. All statistical tests were two-tailed, with significance defined at $p < 0.05$. Analyses were performed in R (version 4.5.1) using the packages data.table (version 1.17.8), GenomicRanges (version 1.62.1), and ggplot2 (version 4.0.2).

3.6.6. Analysis of Associations between Long Non-coding RNAs and Transposable Elements

The association between TE content and lncRNAs, assembly-derived PCGs, and BRAKER3-derived PCGs was assessed by calculating the proportion of total gene, exon, and intron lengths overlapped by major TE classes. Prior to overlap analysis, genomic coordinates for each TE class were collapsed into non-overlapping ranges. Overlap analyses and range collapsing were performed in a strand-dependent manner. The following TE classes were analysed: LTR retrotransposons, DNA transposons, LINEs, SINEs, and helitrons.

To assess whether TE classes were enriched within lncRNA-associated genomic regions relative to genomic background, enrichment analyses were performed for lncRNA genes, exons, and introns. For each genomic element, TE density was defined as the proportion of total element length occupied by a TE class. Genome-wide TE density for each TE class was calculated as the total genomic span occupied by the TE class divided by total decontaminated genome assembly length. Enrichment values were then calculated as the ratio between element TE density and genome-wide TE density. Analyses were performed in R (version 4.5.1) with the following packages: data.table (version 1.17.8), GenomicRanges (version 1.62.1), and ggplot2 (version 4.0.2).

3.6.7. *Evolutionary Conservation Analysis of Long Non-coding RNAs*

Pairwise species sequence similarity was assessed using BLASTn (version 2.2.31) with the dc-megablast mode. BLASTn identifies regions of local nucleotide sequence similarity by extending exact word matches into high-scoring alignments, enabling the detection of potentially homologous sequences (Altschul et al., 1990). Searches were restricted to the forward strand (-strand plus). Hits were retained if they exhibited an E-value lower than 10^{-20} , $\geq 60\%$ sequence identity, an alignment length ≥ 100 bp, and coverage of at least 50% of the shorter sequence. For each query lncRNA, only the hit with the highest bit score was retained for each target species. To assess sequence-level conservation among lncRNA repertoires across the 14 sponge species, two complementary pairwise metrics were derived from the filtered BLASTn results. The first metric quantified the proportion of lncRNAs in the query species with at least one qualifying hit in the target species. Because this measure is directional, pairwise comparisons were symmetrised by averaging reciprocal species comparisons. The resulting similarity matrix was visualised as a heatmap, and species were grouped by hierarchical clustering using complete linkage and Euclidean distance based on these pairwise proportions. For species pairs exhibiting detectable conservation, a second metric was calculated as the product of sequence identity and the minimum of query and subject coverage proportions, thereby penalising both low-identity and short partial alignments. This metric was used to further characterise the strength of sequence conservation among shared lncRNA homologues.

To identify groups of conserved lncRNAs, a sequence similarity network was constructed from the filtered BLASTn results, with lncRNAs represented as nodes and sequence similarities represented as edges. Connected components were identified using the R package igraph (version 2.1.4) and treated as sequence-conserved clusters. Clusters spanning at least three species were retained for downstream analyses. Representative multiple sequence alignments were generated using MAFFT (version 7.526) (Katoh & Standley, 2013) to assess sequence conservation within clusters. To assess whether particular lncRNA classes were preferentially represented within conserved clusters, enrichment analyses were performed using Fisher's exact test, comparing the frequency of each lncRNA class within clusters to its frequency in the analysed genome.

Syntenic relationships between PCG loci across all 14 sponge species were identified using MCScanX (version 1.0.0) (Wang et al., 2012). Protein sequences predicted by BRAKER3 were compared in all pairwise species combinations using DIAMOND BLASTp with the most sensitive alignment mode and an E-value threshold of 10^{-20} (--more-sensitive --evalue 1e-20). Orthologous PCGs were inferred from eggNOG-mapper results. Protein homology results and gene coordinates were used as input for MCScanX with default parameters to identify collinear genomic blocks. To identify syntenic lncRNAs, the three nearest upstream and three nearest downstream PCGs were identified for each lncRNA. A lncRNA was considered to have syntenic support in a partner species if at least two of these neighbouring genes were located

within the same MCScanX-defined syntenic block. Candidate syntenic lncRNA pairs were subsequently filtered using a definition requiring conservation of both lncRNA genomic class and local gene context. Specifically, antisense, intronic, and overlapping lncRNAs were required to be associated with the orthologous PCGs within syntenic neighbourhoods, whereas intergenic and bidirectional lncRNAs were required to share orthologous flanking PCGs both upstream and downstream within a syntenic block. For intergenic and bidirectional lncRNAs, strand-reversed configurations were permitted to account for local genomic inversions. Synteny conservation breadth for each lncRNA was defined as the number of species containing a syntenic counterpart. Pairwise synteny conservation was visualised as an asymmetric matrix in which each cell represents the proportion of lncRNAs in the query species with at least one syntenic counterpart in the subject species. Species were subsequently grouped by hierarchical clustering using complete linkage and Euclidean distance calculated from the pairwise synteny conservation values.

To visualise large-scale genomic synteny between the species pair exhibiting the highest degree of lncRNA conservation, *E. muelleri* and *S. lacustris*, chromosome-level synteny plots were generated. The 15 longest *E. muelleri* chromosomes were selected as reference chromosomes, and for each, the *S. lacustris* chromosome containing the greatest number of shared syntenic blocks was designated as its corresponding partner chromosome. Two overlapping density tracks were plotted for each chromosome pair: a track representing the genomic density of all syntenic lncRNAs and a track depicting the subset of syntenic lncRNAs that exhibited direct sequence similarity to their syntenic counterpart.

To characterise the genomic context of multispecies-conserved lncRNAs, syntenic lncRNAs were grouped into conserved neighbourhoods based on their co-occurrence within shared syntenic blocks across multiple species. Functional annotations of PCGs were used to define the conserved gene context associated with each neighbourhood. The genomic class composition of lncRNAs and the proportion exhibiting sequence conservation were calculated for each conserved neighbourhood.

To systematically characterize evolutionary conservation among sponge lncRNAs, each lncRNA was assigned to one of three conservation categories. Specifically, lncRNAs were classified as sequence-conserved if they exhibited qualifying BLASTn matches to lncRNAs in at least two additional species. Synteny-conserved lncRNAs were defined as those possessing syntenic counterparts in at least two additional species. A third category, sequence- and synteny-conserved, comprised lncRNAs meeting both preceding criteria and additionally exhibiting direct BLASTn matches to all syntenic counterparts.

To investigate the potential biological functions associated with conserved lncRNAs, Gene Ontology (GO) enrichment analysis was performed using GO annotations assigned to PCGs by eggNOG-mapper. The enrichment analysis was performed using topGO (version 2.1.4) (Alexa & Rahnenfuhrer, 2025). Protein-coding genes were assigned to lncRNAs in a class-aware

manner consistent with the synteny definition: the directly overlapping gene was used for antisense, intronic, and overlapping lncRNAs, whereas the immediate upstream and downstream flanking genes were used for intergenic and bidirectional lncRNAs, with the distance threshold set to 10 kb. Enrichment was tested separately for each conservation category and species across the Biological Process and Molecular Function ontologies using Fisher's exact test. All annotated PCGs within each species were used as the statistical background, and GO terms supported by fewer than five annotated genes were excluded. Conservation analysis was performed using R (version 4.5.2) with the following packages used for data manipulation and visualization: data.table (version 1.17.8), GenomicRanges (version 1.62.1), ggplot2 (version 4.0.2), and pheatmap (version 1.0.13).

3.6.8. Long Non-coding RNA Expression Analysis

For expression analysis, trimmed and corrected RNA-seq reads were aligned to the corresponding decontaminated genomes using HISAT2 with strand-specific alignment and a maximum intron length of 10 kb (`--rna-strandness RF --max-intronlen 10000`). Gene annotation files for lncRNAs were manually generated from genomic coordinates, whereas annotations for PCGs were obtained from the BRAKER3 gene predictions. Read counts were quantified using featureCounts (version 2.0.3) (Liao et al., 2014) by counting paired-end reads overlapping annotated exons, aggregating counts at the gene level, and accounting for strand-specific library orientation (`-t exon -g gene_id -p -s 2`). Next, TPM values were calculated by first normalizing read counts by transcript length to obtain reads per kilobase, followed by scaling to the total number of mapped reads in each library and multiplying by one million. Expression levels were compared between expressed (TPM > 0) lncRNA genes and PCGs annotated by BRAKER3 using a two-tailed Mann–Whitney U test. P-values were adjusted for multiple testing using the Benjamini–Hochberg procedure, with statistical significance defined as $p < 0.05$. Differences in expression among lncRNA classes, defined according to their genomic relationship to PCGs, were assessed using a two-tailed Kruskal–Wallis test with Benjamini–Hochberg correction ($p < 0.05$). Finally, sample-specific lncRNA expression patterns were evaluated based on the number of distinct libraries in which they were expressed (TPM > 0) per species. All analyses were performed in R (version 4.5.1) using the packages data.table (version 1.17.8) and ggplot2 (version 4.0.2).

3.6.9. Differential Expression Analysis of Long Non-Coding RNAs in Response to Microbial-Associated Molecular Patterns

RNA-seq libraries generated by Pita et al. (2018) were reanalysed to investigate lncRNA responses to microbial-associated molecular pattern (MAMP) exposure in the marine demosponges *A. aerophoba* and *D. avara*. The dataset comprised 35 libraries representing control and MAMP-treated individuals sampled at 1, 3, and 5 h post-exposure. Three biological replicates were available for each treatment and time-point combination, with the exception of the *D. avara* MAMP treatment at 1 h post-exposure, for which only two replicates were

available. These libraries represented a subset of the transcriptomic datasets used for lncRNA characterization in both species. Read count matrices generated during the lncRNA expression quantification workflow (Section 3.6.8) were reused for all downstream differential expression and co-expression analyses.

Prior to analysis, lowly expressed genes were removed by retaining only genes with at least 10 reads in a minimum of three libraries. Variance-stabilising transformation was applied for quality control and visualisation, and PCA was performed on variance-stabilised expression values to assess sample-level structure and identify potential outliers. Raw read counts for the stated libraries were analysed using DESeq2 (version 1.53.0) (Love et al., 2014). A single design factor encoding the combination of treatment and time point was used to model gene expression, allowing independent contrasts between MAMP-treated and control samples at each time point. Differential expression was assessed using two-tailed Wald tests with Benjamini–Hochberg correction for multiple testing. Log₂ fold-change shrinkage was performed using the adaptive shrinkage estimator implemented in the ashR package (version 2.2.63) (Stephens et al., 2023). Genes were considered differentially expressed when the adjusted p-value was < 0.05 and the absolute fold change exceeded 1.5. Differential expression results were visualised using volcano plots, with lncRNA genes and PCGs plotted separately. To visualise expression patterns among differentially expressed genes (DEGs), variance-stabilised expression values were scaled by gene (row z-score) and subjected to hierarchical clustering using complete linkage and Euclidean distance.

Weighted gene co-expression network analysis (WGCNA) was performed separately for each species using the WGCNA package (version 1.74) (Langfelder & Horvath, 2008). Expression matrices were constructed from variance-stabilised expression values of all genes passing the expression filter, after removal of the bottom 25% of genes ranked by expression variance. Gene and sample quality were assessed using the goodSamplesGenes function. Soft-thresholding powers were selected independently for each species as the lowest power producing a scale-free topology fit of $R^2 > 0.80$. Signed hybrid networks were constructed using blockwiseModules with a signed topological overlap matrix, with minimum module size of 30 genes, and module merging threshold of 0.40. Soft-thresholding powers of 14 and 9 were selected for *A. aerophoba* and *D. avara*, respectively. Module eigengenes, representing the first principal component of expression within each module, were correlated with binary variables encoding MAMP treatment and treatment-by-time-point interactions. Correlations were calculated using Pearson correlation, and significance was assessed using the Student's t-test approximation implemented in corPvalueStudent function. Modules were retained for downstream analysis if they met all of the following criteria: (i) significant correlation between the module eigengene and at least one MAMP-associated trait ($p < 0.05$), (ii) presence of at least one lncRNA-PCG co-expression pair with Pearson correlation coefficient > 0.85 , and (iii) a minimum of 20 lncRNAs and 20 PCGs within the module. The grey module was excluded, as it represents genes not assigned to any co-expression cluster.

Module eigengene distributions across treatments and time points were visualised using boxplots. Next, GO enrichment analysis was performed on PCGs assigned to each retained module using the `enricher` function from `clusterProfiler` (version 4.20.0) (Wu et al., 2021). Functional annotations were derived from `eggNOG-mapper` output. Enrichment was assessed against the complete set of annotated PCGs expressed in the respective species, considering Biological Process ontology terms only. Gene sets containing between 5 and 500 genes were retained, and enriched terms were considered significant at the adjusted p-value threshold of 0.05.

To characterise co-expression relationships within modules, pairwise Pearson correlations were calculated between all lncRNA and PCG expression profiles. Pairs with Pearson correlation coefficients > 0.80 were retained as candidate co-expression relationships. Network visualisations were generated using `igraph` (version 2.1.4) and `ggraph` (version 2.2.2) (Pedersen, 2025), with edges restricted to correlations > 0.80 . To improve visual interpretability, 100 highest-confidence edges per module were displayed. Node size was scaled according to the number of connections, and node positions were determined using the Fruchterman–Reingold force-directed layout algorithm. Hub lncRNAs were defined as transcripts occupying the upper 10% of the number of connections within each module network and were characterised according to network connectivity, module membership, differential expression status, and genomic classification.

Differential expression analysis and WGCNA were performed in R (version 4.5.2), with the following packages used for data manipulation and visualization: `data.table` (version 1.17.8), `ggplot2` (version 4.0.2), `pheatmap` (version 1.0.13), and `cowplot` (version 1.1.3).

4. RESULTS

4.1. CATS Benchmarking

To address the limitations of existing transcriptome assembly quality assessment methods, this dissertation developed CATS, a novel framework consisting of two complementary modules. CATS-rf is a reference-free algorithm which evaluates transcript-level accuracy based on RNA-seq read support, while CATS-rb represents a reference-based algorithm that assesses transcriptome structural accuracy and completeness through comparative transcript-to-genome mapping.

The performance of both CATS modules was systematically benchmarked using multiple complementary datasets. Controlled simulations followed an orthogonal experimental design in which transcript coverage, sequencing error rate, and assembler choice were systematically varied across six model species. This framework allowed the individual effects of each variable on CATS scoring to be assessed independently.

Realistic simulations incorporated expected intra-library variation in transcript coverage, along with biologically plausible proportions of expressed transcripts. Within this framework, mean transcript coverage and assembler choice were varied across the same six species, with each simulation replicated four times. This dataset was further complemented by publicly available RNA-seq libraries for the same six species, enabling evaluation in realistic experimental settings.

Additionally, artificially mutated transcriptomes were generated by introducing five distinct mutation types across four progressively increasing severity levels, as well as through cumulative combinations of all mutation types across six severity levels. This design enabled detailed assessment of CATS scoring in response to increasing levels of most common assembly errors.

4.1.1. Association of CATS-rf Scores with Assembly Features and Quality

CATS-rf was initially evaluated on the controlled simulated dataset, which comprised 504 transcriptome assemblies. Transcript score distributions differed across the analysed species and assemblers (Figure 13). More complex transcriptomes, including those of *M. musculus* and *H. sapiens*, generally produced lower transcript scores. Across all species, rnaSPAdes and Trinity consistently achieved superior performance compared with IDBA-tran and SOAPdenovo-Trans.

Low coverage levels resulted in substantially lower transcript scores across all assemblies (Figure 14). In most cases, increasing coverage beyond 5–10× provided only marginal improvements, although these gains were more pronounced in libraries with elevated

sequencing error rates. Furthermore, higher sequencing error rates were associated with reduced transcript scores across all coverage levels.

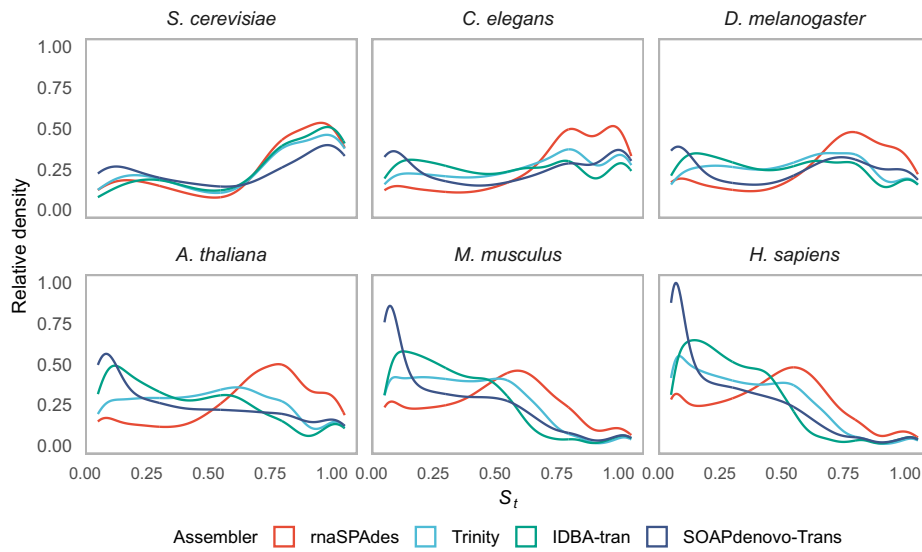


Figure 13. Distribution of CATS-rf transcript scores across species and assemblers in controlled simulations. S_t = CATS-rf transcript score. Reproduced from Bodulić & Vlahoviček, 2026.

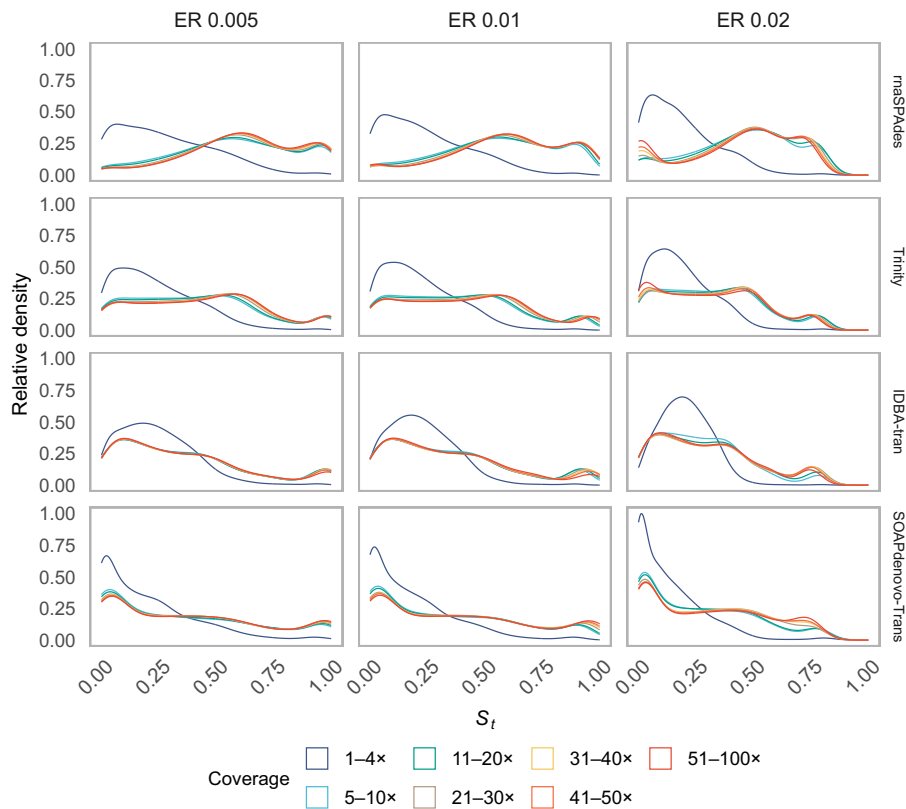


Figure 14. Distribution of CATS-rf transcript scores across assemblers, sequencing error rates, and coverage levels in controlled simulations. ER = error rate, S_t = CATS-rf transcript score. Reproduced from Bodulić & Vlahoviček, 2026.

To assess redundancy among the CATS-rf score components, pairwise correlations were calculated within each assembly and subsequently averaged across all assemblies. Weak to moderate intercorrelations were observed among the score components, with median ρ values ranging from 0.02 (IQR = 0.24) to 0.46 (IQR = 0.20) (Figure 15).

S_i	0.2	0.46	0.21	1
S_l	0.21	0.22	1	0.21
S_a	0.02	1	0.22	0.46
S_c	1	0.02	0.21	0.2
	S_c	S_a	S_l	S_i

Figure 15. Median Spearman correlation coefficients between CATS-rf score components across assemblies from controlled simulations. S_c = coverage component, S_a = accuracy component, S_l = local fidelity component, S_i = integrity component. Reproduced from Bodulić & Vlahoviček, 2026.

Pairwise correlations were further examined within assembly subsets stratified by selected species, coverage range, sequencing error rate, and assembler. Correlation coefficients were consistently low across subsets (Figure 16). The most pronounced differences were observed between assemblies with differing coverage levels. For instance, in low-coverage assemblies (1–4×), a negative correlation was observed between *the* coverage and accuracy *component* (median ρ = −0.22, IQR = 0.27). Additionally, a notable discrepancy was identified in the correlation between the accuracy and integrity components, which increased from median ρ = 0.11 (IQR = 0.20) at 1–4× coverage to median ρ = 0.49 (IQR = 0.18) at ≥5–10× coverage.

The performance of CATS-rf was first evaluated against existing reference-free transcriptome assessment tools using the controlled dataset, allowing performance to be assessed under idealized conditions spanning a broad range of assembly characteristics. Transcript scores from CATS-rf, RSEM-EVAL, and TransRate were correlated to transcript F-scores derived from reciprocal best hits to reference transcripts, which served as an objective measure of transcript accuracy. These correlations were initially calculated within each assembly. Among the evaluated methods, CATS-rf transcript scores demonstrated the strongest association with F-scores (median ρ = 0.75, IQR = 0.17), substantially outperforming both RSEM-EVAL (median ρ = 0.38, IQR = 0.51) and TransRate (median ρ = 0.46, IQR = 0.23) (Figure 17). The CATS-rf coverage component also showed a stronger correlation with F-scores (median ρ = 0.42, IQR = 0.37) than the corresponding TransRate metric (median ρ = 0.16, IQR = 0.18). Likewise, the CATS-rf accuracy component (median ρ = 0.42, IQR = 0.30) outperformed its TransRate equivalent (median ρ = −0.15, IQR = 0.28). Whereas TransRate integrates evidence of

structural errors and fragmentation into a single “order” score, CATS-rf distinguishes these features into separate components, both of which exhibited moderate correlations with F-scores (local fidelity: median $\rho = 0.38$, IQR = 0.40; integrity: median $\rho = 0.63$, IQR = 0.23). While CATS-rf does not include a dedicated chimerism detection component, the segmentation score implemented in TransRate showed only weak correlation with transcript F-scores (median $\rho = 0.19$, IQR = 0.24).

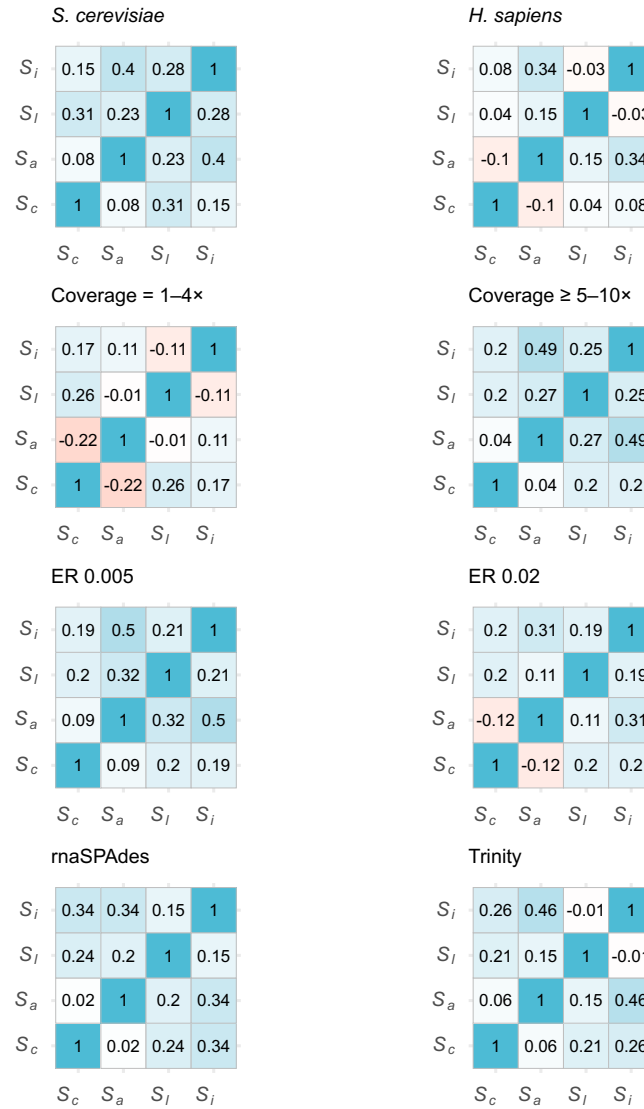


Figure 16. Median Spearman correlation coefficients between CATS-rf score components across assemblies stratified by selected species, coverage levels, sequencing error rates, and assemblers. S_c = coverage component, S_a = accuracy component, S_l = local fidelity component, S_i = integrity component, ER = error rate. Reproduced from Bodulić & Vlahoviček, 2026.

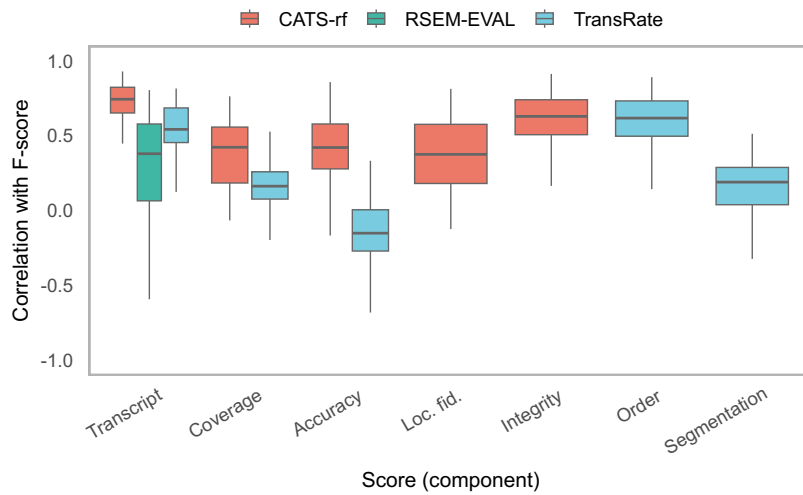


Figure 17. Distribution of Spearman correlation coefficients between transcript F-scores and transcript scores from CATS-rf, RSEM-EVAL, and TransRate across assemblies from controlled simulations. Boxplots represent the median and IQR, with whiskers extending to $\pm 1.5 \times \text{IQR}$. Reproduced from Bodulić & Vlahoviček, 2026.

Stratified analyses revealed that these performance differences were consistent across species, coverage ranges, sequencing error rates, and assemblers, with CATS-rf demonstrating the highest accuracy across the vast majority of assembly subsets (Figure 18). The greatest performance advantage was observed in less complex transcriptomes (*S. cerevisiae*) and among higher-coverage transcripts, whereas sequencing error rate and assembler choice had comparatively weaker effects on the overall performance. In contrast, RSEM-EVAL showed a marked decline in performance for rnaSPAdes assemblies. Notably, the CATS-rf coverage and local fidelity score components displayed lower correlations with F-scores in more complex transcriptomes (*H. sapiens*), while the accuracy component showed weaker performance for low-coverage transcripts. The accuracy component also exhibited weaker performance under higher sequencing error conditions (0.02), while both the accuracy and local fidelity components showed slightly reduced performance in Trinity assemblies. The fragmentation component was relatively stable across all analysed subsets.

A more detailed analysis across species, assemblers, and coverage levels revealed that CATS-rf consistently outperformed TransRate across all benchmarking conditions (Figure 19). Similar patterns were observed when comparing CATS-rf with RSEM-EVAL: CATS-rf achieved superior performance across nearly all conditions, with the exception of the lowest coverage levels, where both methods performed comparably. In higher-coverage assemblies, however, RSEM-EVAL exhibited substantial inaccuracy, particularly in less complex transcriptomes (*S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *A. thaliana*).

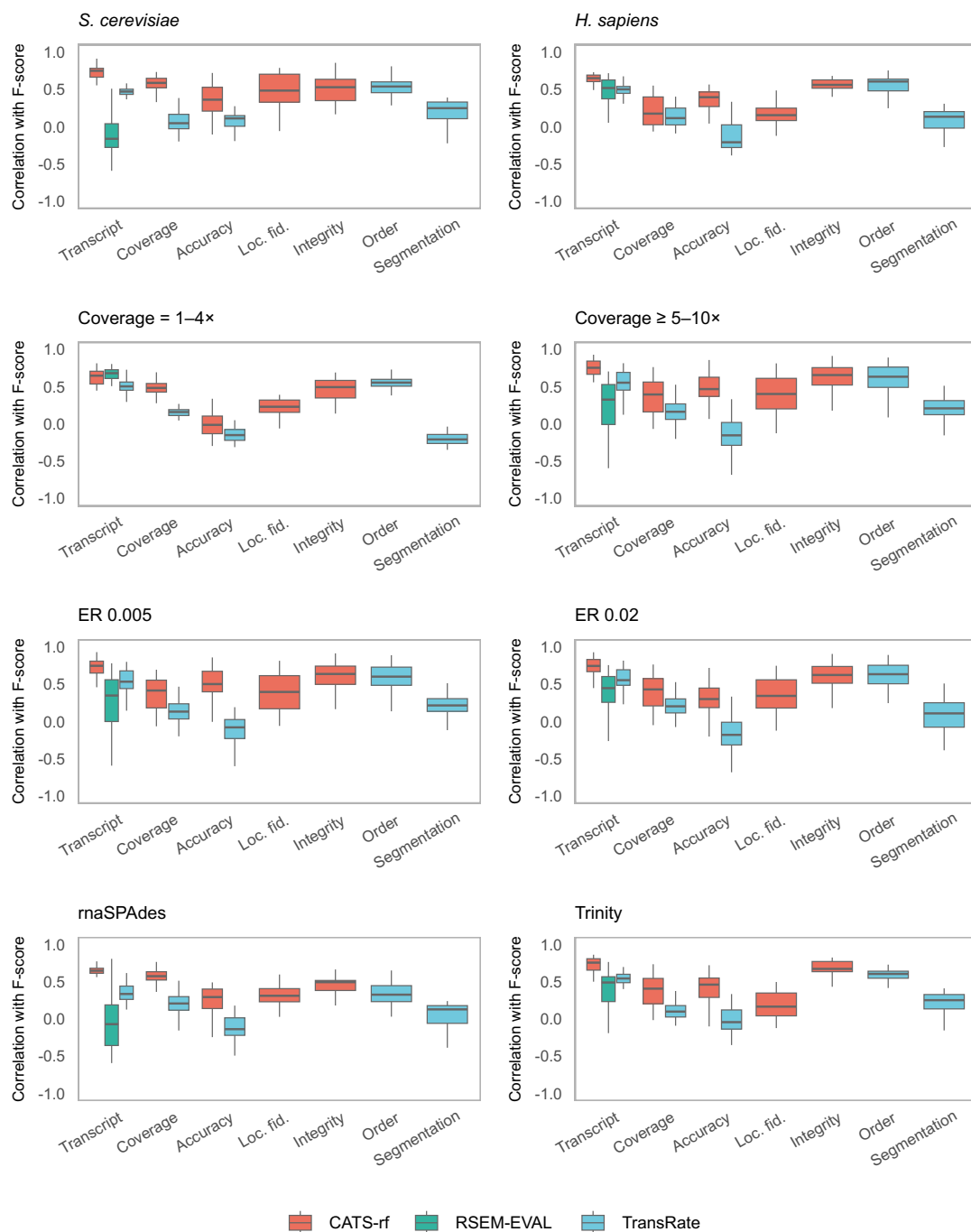


Figure 18. Distribution of Spearman correlation coefficients between transcript F-scores and transcript scores from CATS-rf, RSEM-EVAL, and TransRate across assemblies stratified by selected species, coverage levels, sequencing error rates, and assemblers. ER = error rate. Reproduced from Bodulić & Vlahoviček, 2026.

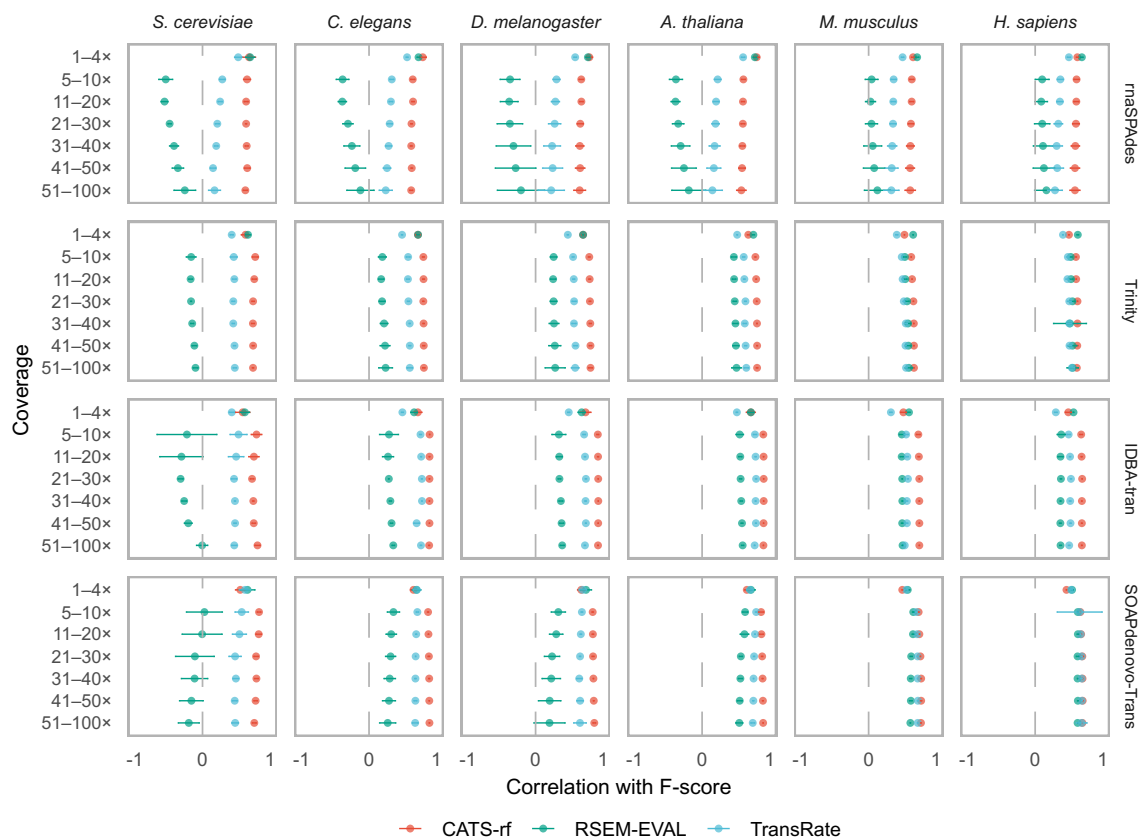


Figure 19. Spearman correlation coefficients of transcript scores from CATS-rf, RSEM-EVAL, and TransRate with transcript F-scores across the analysed species, coverage levels, and assemblers. Each point represents the median across $N = 3$ assemblies corresponding to libraries with distinct sequencing error rates. Error bars indicate the IQR. Reproduced from Bodulić & Vlahoviček, 2026.

The previous evaluations focused primarily on how well each method correlated with relative assembly performance within individual assemblies. Here, this assessment is extended to determine each tool's ability to rank datasets based on absolute assembly accuracy across heterogeneous library conditions, providing a more rigorous test across diverse experimental contexts. This was done by evaluating the correlation between mean transcript quality scores produced by each tool and mean transcript F-scores. Again, CATS-rf demonstrated superior performance, exhibiting a significantly stronger correlation with mean F-scores ($\rho = 0.97$, 95% CI 0.96–0.97) than both RSEM-EVAL ($\rho = 0.40$, 95% CI 0.32–0.47, $z = 28.25$, $p < 0.001$) and TransRate ($\rho = 0.88$, 95% CI 0.88–0.88, $z = 14.24$, $p < 0.001$) (Figure 20).

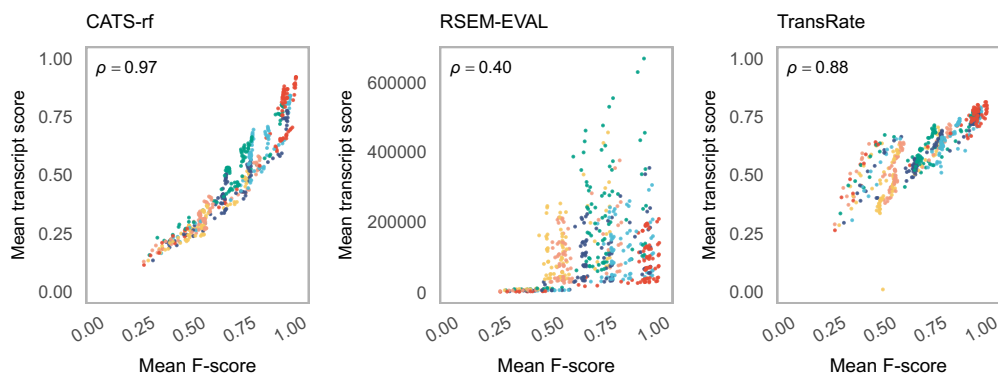


Figure 20. Correlation of mean transcript scores from CATS-rf, RSEM-EVAL, and TransRate with mean transcript F-scores across assemblies from controlled simulations. Reproduced from Bodulić & Vlahoviček, 2026.

CATS-rf was subsequently evaluated using 384 realistically simulated transcriptome assemblies. Transcript score distributions exhibited profiles intermediate between low- and high-coverage libraries (Figure 21). CATS-rf scores remained positively associated with coverage, while more complex transcriptomes generally displayed lower overall quality. The strongest performance was again observed for rnaSPAdes and Trinity.

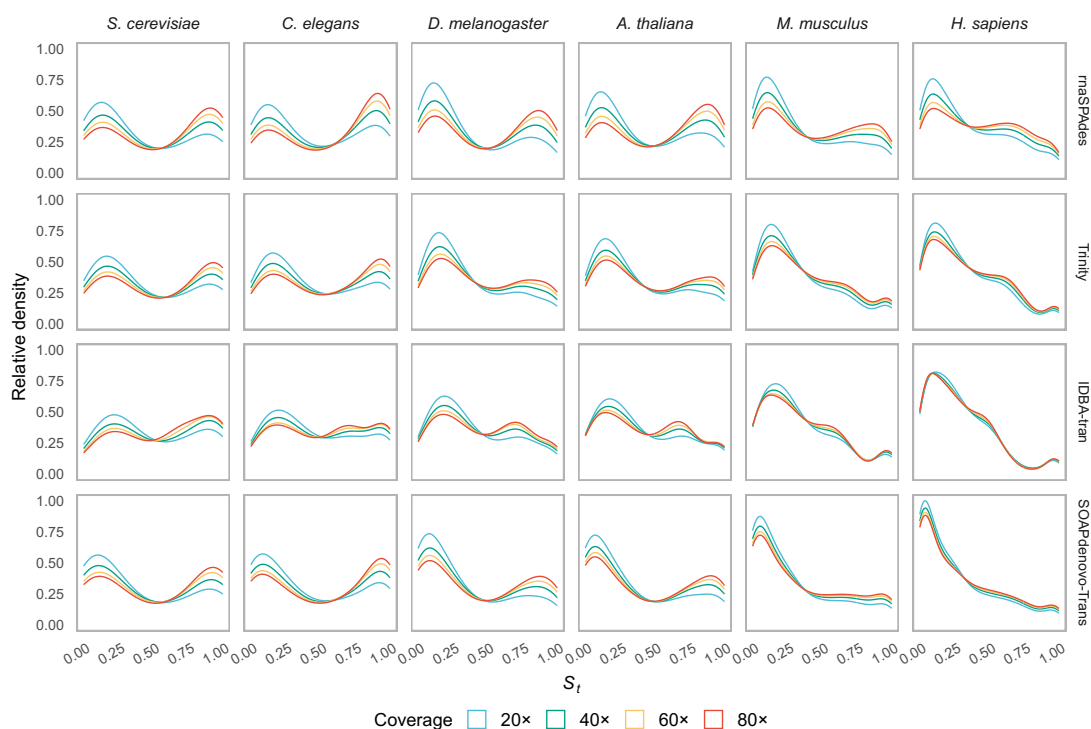


Figure 21. Distribution of relative transcript scores across assemblers and coverage in realistically simulated libraries. Reproduced from Bodulić & Vlahoviček, 2026.

CATS-rf showed substantially stronger within-assembly correlations between transcript scores and F-scores than either RSEM-EVAL or TransRate, with a median ρ of 0.85 (IQR = 0.10),

compared with 0.73 (IQR = 0.12) for RSEM-EVAL and 0.57 (IQR = 0.13) for TransRate. It also outperformed both methods across the vast majority of datasets, with the exception of six low-coverage SOAPdenovo-Trans assemblies (Figure 22). Genrally, all tools exhibited similar accuracy in SOAPdenovo-Trans assemblies.

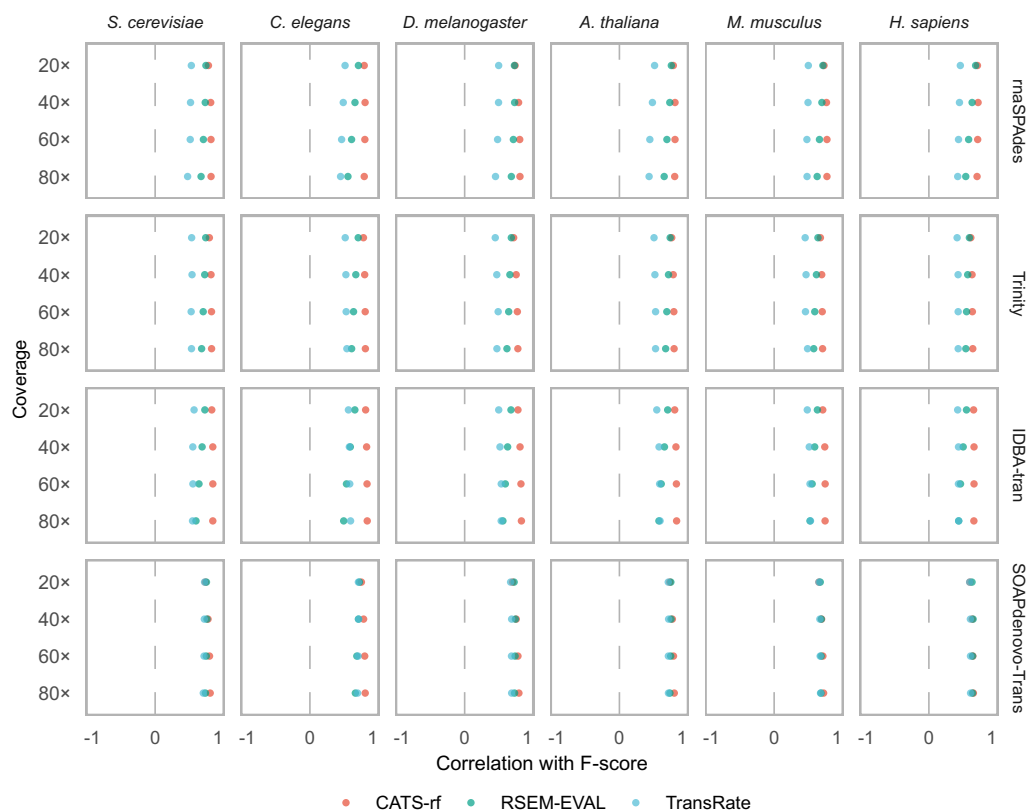


Figure 22. Spearman correlation coefficients of transcript scores from CATS-rf, RSEM-EVAL, and TransRate with transcript F-scores across assemblies from realistically simulated libraries. Each point represents the median across $N = 4$ library replicates for each species and coverage combination. Error bars are omitted due to minimal variability across replicates (IQR < 0.03). Reproduced from Bodulić & Vlahoviček, 2026.

At the assembly level, mean CATS-rf transcript scores were more strongly associated with mean transcript F-scores ($p = 0.93$, 95% CI 0.92–0.95) than transcript accuracy estimates from RSEM-EVAL ($p = 0.31$, 95% CI 0.21–0.40, $z = 20.17$, $p < 0.001$) or TransRate ($p = 0.78$, 95% CI 0.74–0.82, $z = 11.11$, $p < 0.001$) (Figure 23).

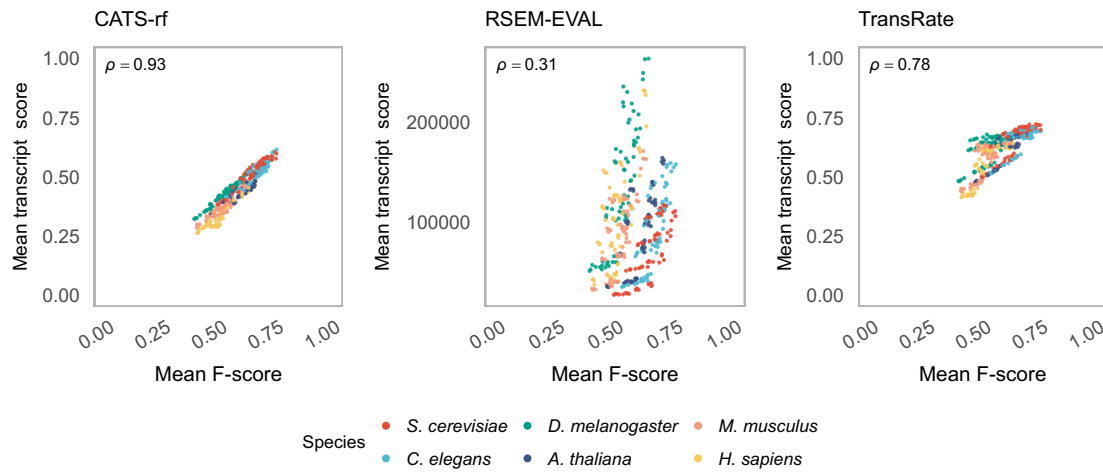


Figure 23. Correlation of mean transcript scores from CATS-rf, RSEM-EVAL, and TransRate with mean transcript F-scores across realistically simulated assemblies. Reproduced from Bodulić & Vlahoviček, 2026.

Finally, CATS-rf was applied to 168 transcriptomes generated from public RNA-seq libraries. Overall, transcript scores assigned by CATS-rf were lower than those observed for simulated assemblies (Figure 24). More complex transcriptomes were generally associated with reduced scores, consistent with patterns observed in simulated datasets. Among the evaluated assemblers, the highest scores were again achieved by rnaSPAdes and Trinity.

Comparative benchmarking was conducted using transcripts represented in the corresponding reference transcriptomes. In this analysis, the strongest correlation between mean transcript scores and mean F-scores was demonstrated by CATS-rf ($\rho = 0.71$, 95% CI 0.62–0.78), exceeding the performance of both RSEM-EVAL ($\rho = 0.35$, 95% CI 0.21–0.47, $z = 5.14$, $p < 0.001$) and TransRate ($\rho = 0.41$, 95% CI 0.28–0.53, $z = 4.51$, $p < 0.001$) (Figure 25).

4.1.2. Association of CATS-rf Scores with Assembly Errors

The ability of CATS-rf to detect transcript errors was evaluated using artificially mutated transcriptome assemblies specifically designed to isolate the impact of distinct mutation types on transcript scoring under conditions of sufficient sequencing coverage. Four levels of mutations were independently introduced into 12 highly covered transcriptome assemblies generated from controlled simulations. The mutation categories included insertions, mismatches, deletions, fragmentation, and redundancy.

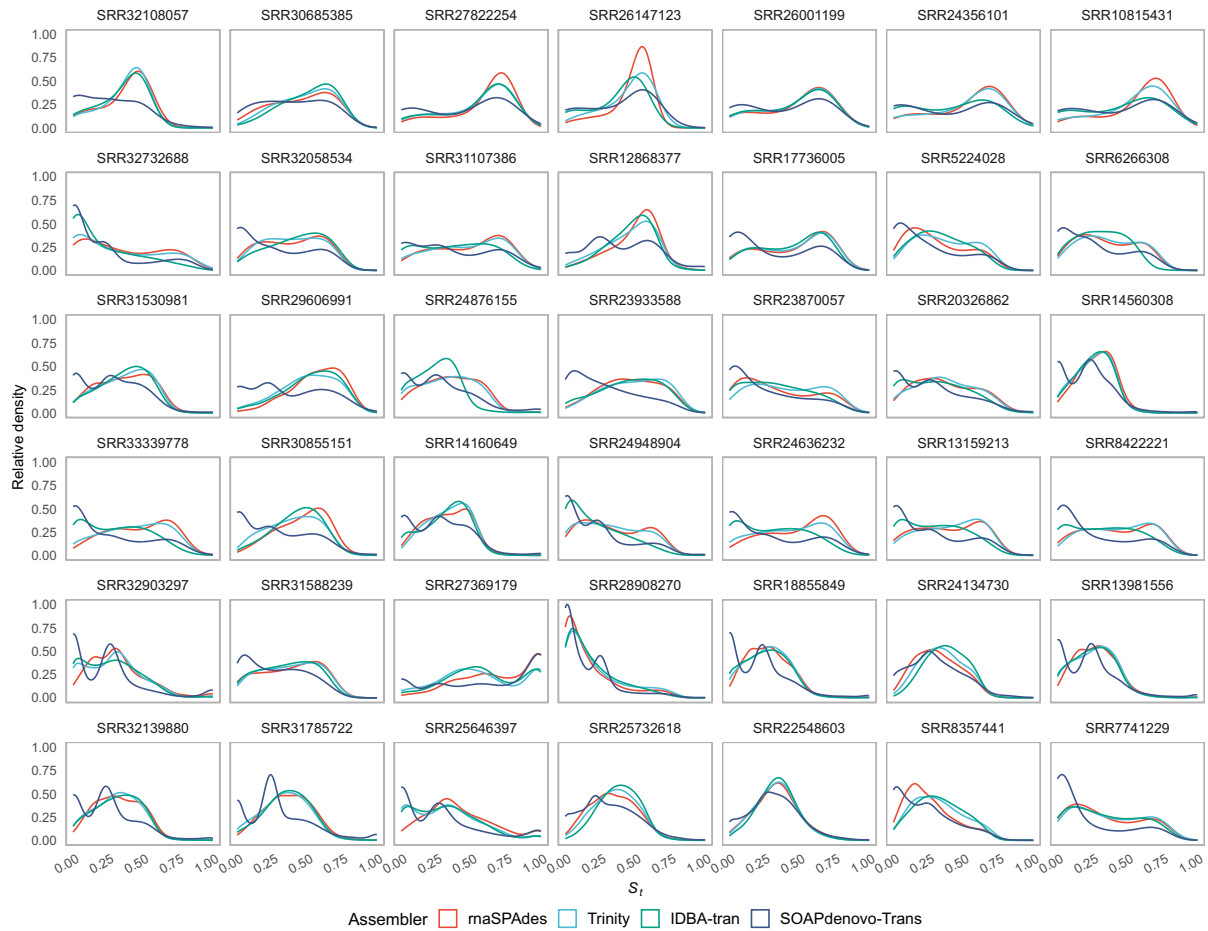


Figure 24. Distribution of CATS-rf transcript scores in assemblies generated from public libraries, labeled by corresponding SRA accession numbers. Reproduced from Bodulić & Vlahoviček, 2026.

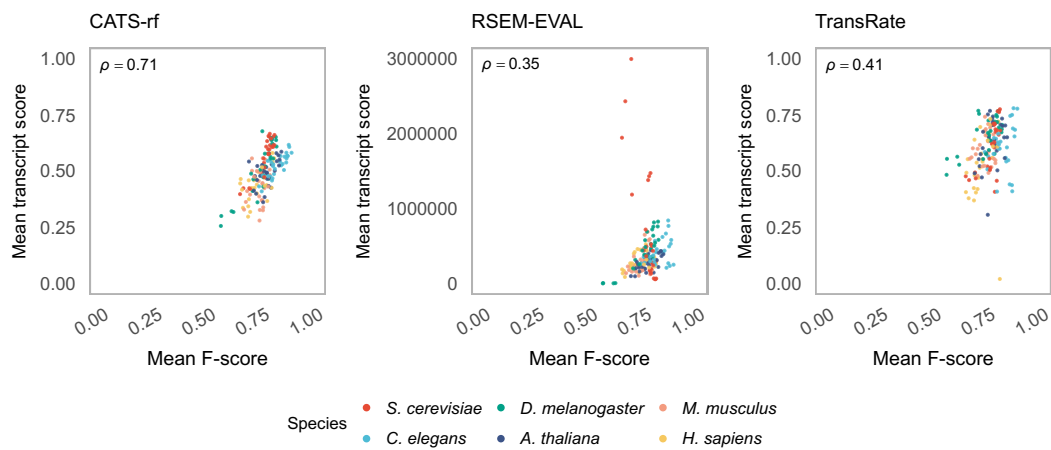


Figure 25. Correlation of mean transcript scores from CATS-rf, RSEM-EVAL, and TransRate with mean transcript F-scores across assemblies from publicly available libraries. Reproduced from Bodulić & Vlahoviček, 2026.

Mutated transcriptome assemblies displayed a substantial decline in CATS-rf assembly score, proportional to the mutation level (Figure 26). An exception was observed in redundant assemblies, where the assembly score plateaued at the third redundancy level (60% of transcripts with 60–79% sequence duplication).

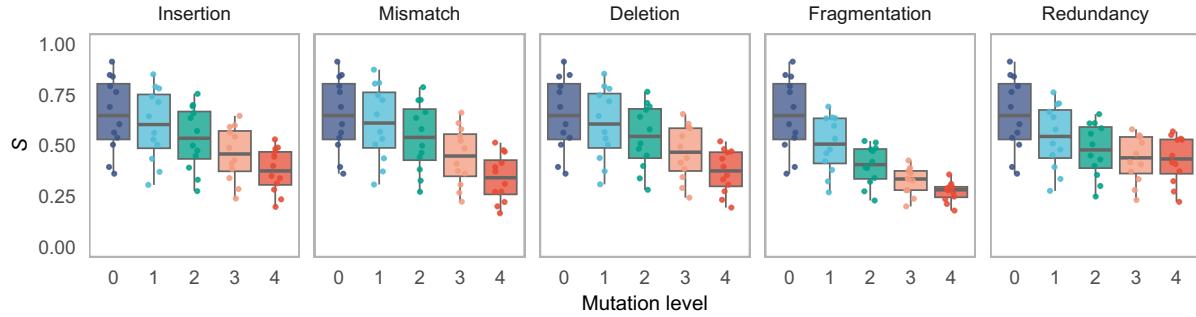


Figure 26. Distribution of CATS-rf assembly scores per mutation level with increasing levels of insertions, mismatches, deletions, fragmentation, and redundancy. Mutation level 0 corresponds to native assemblies. S = assembly score. Reproduced from Bodulić & Vlahoviček, 2026.

Coverage, accuracy, and local fidelity score components were also analysed across transcripts containing progressively higher levels of insertions, mismatches, and deletions, respectively. Significant reductions were observed across all corresponding score components for each mutation type (Figure 27). The strongest decline was found in the coverage component of transcripts with insertions.

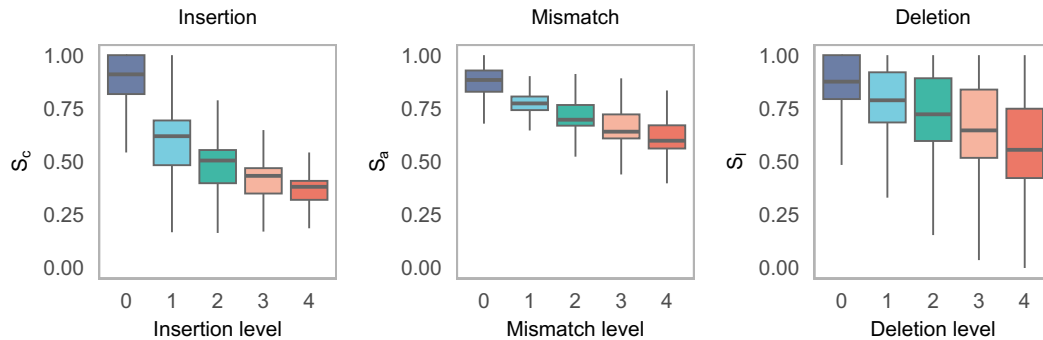


Figure 27. Distribution of CATS-rf coverage, accuracy, and local fidelity score components in transcripts with increasing levels of insertions, mismatches, and deletions, respectively. Mutation level 0 corresponds to native assemblies. S_c = coverage component, S_a = accuracy component, S_l = local fidelity component. Reproduced from Bodulić & Vlahoviček, 2026.

Similarly, fragmented assemblies exhibited a significant reduction in the integrity component (Figure 28).

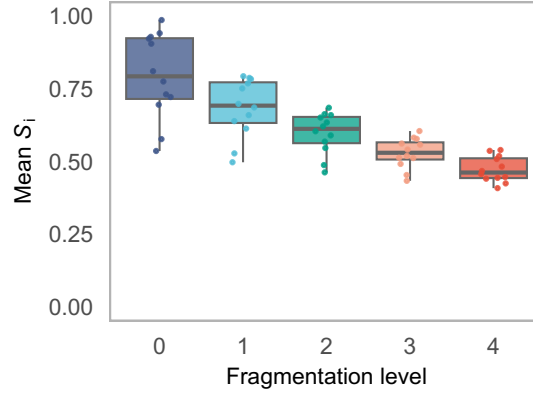


Figure 28. Distribution of mean CATS-rf integrity score component across assemblies with increasing fragmentation. Mutation level 0 corresponds to native assemblies. S_i = integrity component. Reproduced from Bodulić & Vlahoviček, 2026.

Redundant assemblies displayed a significant reduction in the coverage component. As with the overall assembly score, this decline became less pronounced after the third redundancy level (Figure 29).

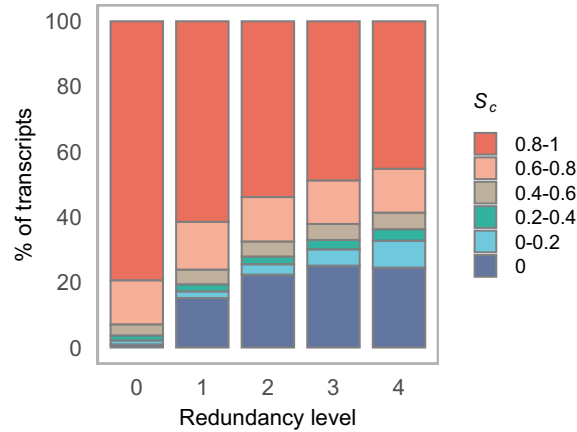


Figure 29. Distribution of the CATS-rf coverage score component across assemblies with increasing redundancy. Mutation level 0 corresponds to native assemblies. S_c = coverage component. Reproduced from Bodulić & Vlahoviček, 2026.

Scoring patterns associated with different mutation types were explored using PCA on mean score components of assemblies with all five mutation types at mutation levels 3 and 4. Distinct clusters emerged based on mutation type, with partial overlap between assemblies containing insertions and redundancy, as well as between assemblies with insertions and mismatches (Figure 30).

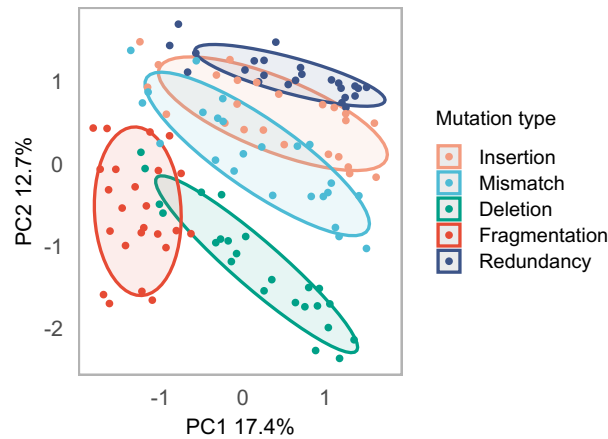


Figure 30. Principal component analysis of mean CATS-rf transcript score components in assemblies with distinct mutation types. Reproduced from Bodulić & Vlahoviček, 2026.

The second part of the artificially mutated dataset comprised assemblies simultaneously subjected to all five mutation types. These assemblies exhibited a significant reduction in CATS-rf assembly scores proportional to the level of mutational burden, with score variance also decreasing substantially (Figure 31).

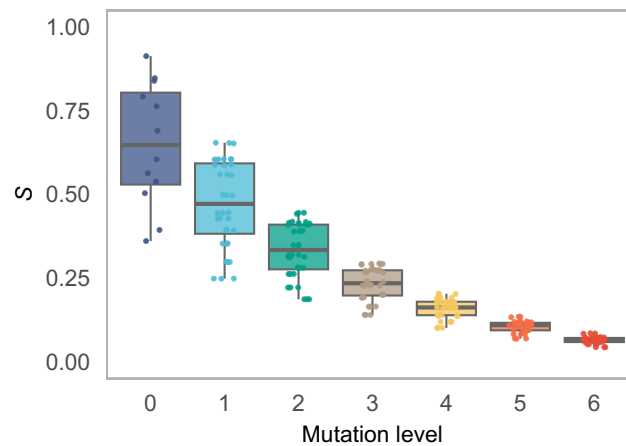


Figure 31. Distribution of assembly scores in transcriptome assemblies with increasing levels of multiplicative mutations. Mutation level 0 corresponds to native assemblies. S = assembly score. Reproduced from Bodulić & Vlahoviček, 2026.

4.1.3. CATS-rf Parameter Robustness Analysis

The effect of CATS-rf parameters on score components was evaluated using a grid-search analysis with randomly sampled parameter sets for each component in 16 realistically simulated assemblies. Across a sensible range of parameter settings, CATS-rf score components exhibited minimal variability (Figure 32). The largest fluctuations were observed in the accuracy and local fidelity components. However, these variations remained relatively minor.

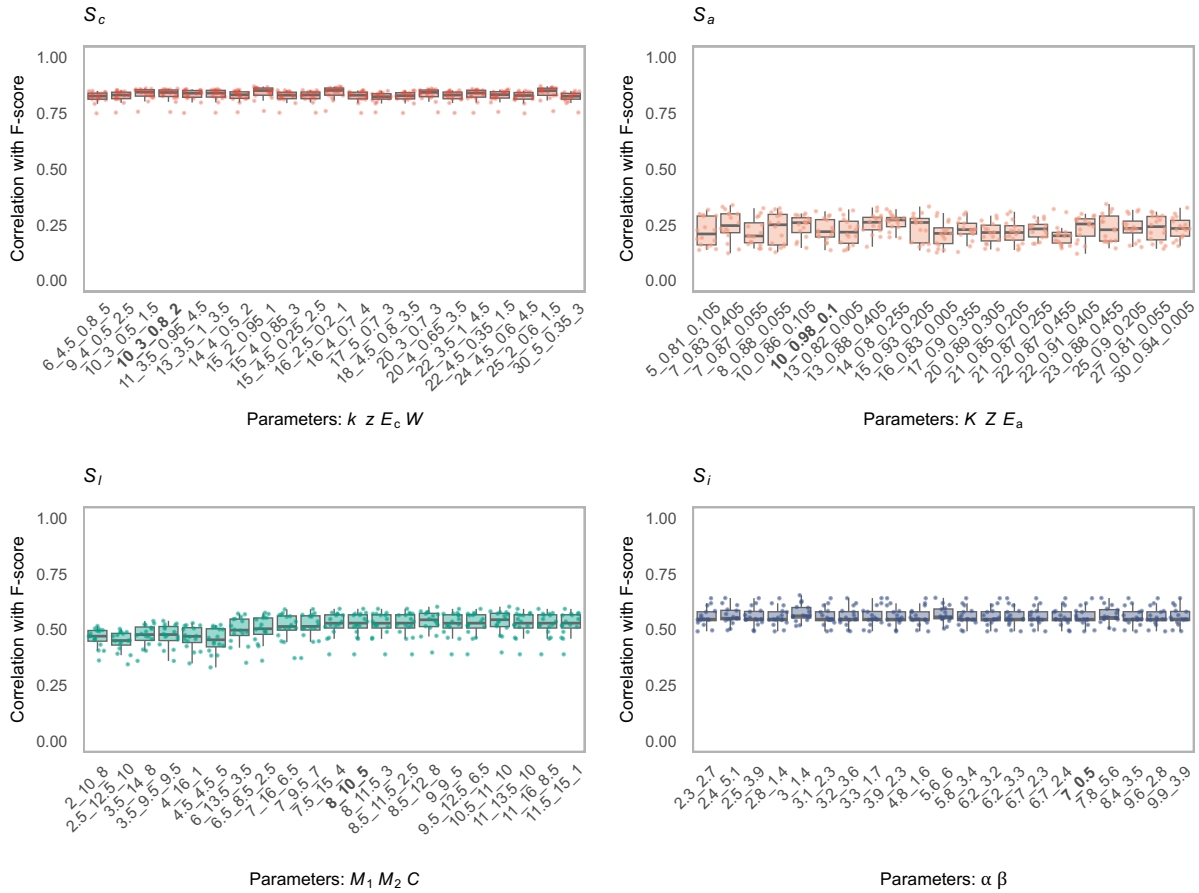


Figure 32. Distribution of correlation coefficients between CATS-rf score components and transcript F-scores across CATS-rf runs with parameter sets sampled via random grid-search (x-axis). Default parameter values are shown in bold. S_c = coverage component, S_a = accuracy component, S_l = local fidelity component, S_i = integrity component, k = rolling window length for LCR detection, z = mean rolling window coverage threshold for LCR definition, E_c = LCR extension penalty, W = base coverage weight, K = rolling window length for LAR detection, Z = mean rolling window accuracy threshold, E_a = LAR extension penalty, M_1 = lower multiplicative factor, M_2 = higher multiplicative factor, C = correction factor, α = compression factor 1, β = compression factor 2. Reproduced from Bodulić & Vlahoviček, 2026.

4.1.4. Association of CATS-rb Scores with Assembly Features and Quality

CATS-rb benchmarking was performed on the same simulated and public datasets utilized in CATS-rf performance evaluation. The benchmark mainly focused on transcript and exon scores, which represent the main completeness metrics provided by CATS-rb. General features of CATS-rb completeness scores were first evaluated on the assemblies from controlled simulations. Relative transcript scores varied substantially across assemblers and species, with complex transcriptomes generally showing weaker performance (Figure 33). Consistent with CATS-rf results, rnaSPAdes and Trinity outperformed IDBA-tran and SOAPdenovo-Trans across all species. Reference assemblies yielded consistently high relative completeness scores (transcript: range 0.95–0.99, exon: range 0.97–0.99).

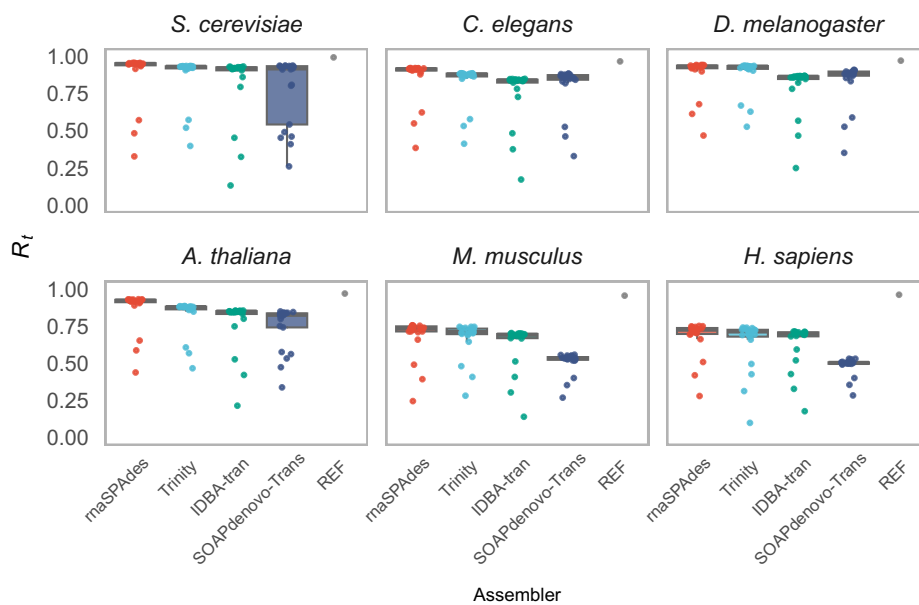


Figure 33. Distribution of CATS-rb relative transcript scores across species and assemblers in controlled simulations. R_t = relative transcript score, REF = reference transcriptome. Reproduced from Bodulić & Vlahoviček, 2026.

CATS-rb transcript scores were also compared across the analysed sequencing error rates and coverage levels. Assemblies generated at the lowest coverage levels exhibited the poorest performance, with transcript scores generally plateauing beyond 5–10× coverage (Figure 34). In contrast, assemblies with the highest sequencing error rates continued to show significant improvement with increasing sequencing depth. Overall, higher error rates were associated with lower transcript scores.

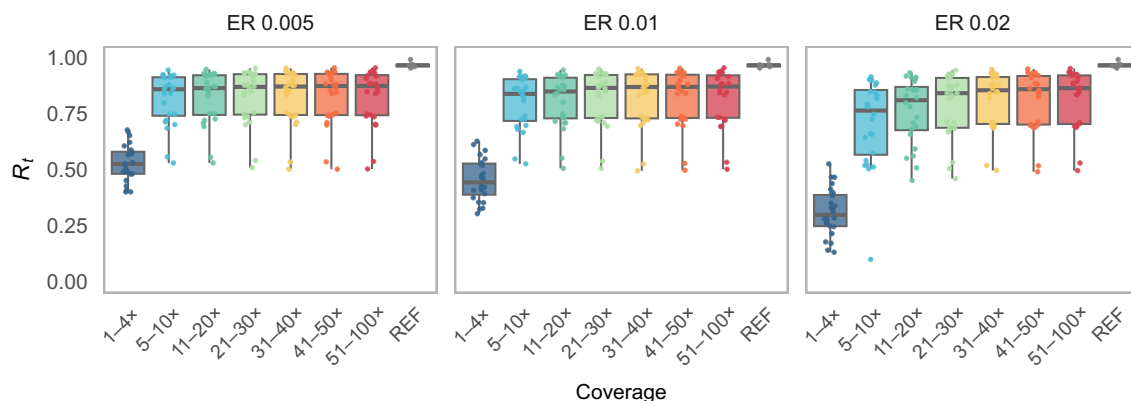


Figure 34. Distribution of CATS-rb relative transcript scores across sequencing error rates and coverage levels in controlled simulations. R_t = relative transcript score, REF = reference transcriptome. Reproduced from Bodulić & Vlahoviček, 2026.

Relative CATS-rb scores were benchmarked against annotation-based CATS-rb scores derived from reference genome annotations to assess their behavior across a broad spectrum of assembly quality. This evaluation was conducted using the complete controlled dataset and five assembly subsets, each representing progressively lower coverage levels. For each species, the complete dataset comprised the reference transcriptome together with all assemblies generated from controlled simulations. In contrast, each subset included only simulated assemblies and was categorized based on the coverage range of its highest-coverage assemblies, spanning from 51–100× down to 1–4×. Relative CATS-rb scores showed a consistently strong correlation with annotation-based scores across all datasets (Figure 35). Notably, this relationship remained highly robust even in the lowest-coverage subset (transcript: $p = 0.94$, 95% CI 0.90–0.96; exon: $p = 0.95$, 95% CI 0.93–0.97).

Relative CATS-rb scores were further analysed using realistically simulated and public datasets. In the realistic simulations, CATS-rb was first applied to the complete per-species assembly sets, including all libraries and the appropriate reference transcriptome. These assemblies generally exhibited lower transcript scores than the assemblies from controlled simulations (Figure 36). Scores increased with greater sequencing depth and were higher for assemblies generated by rnaSPAdes and Trinity. Reference transcriptomes again exhibited near-maximum scores (transcript: range 0.97–0.99, exon: range 0.98–0.99).

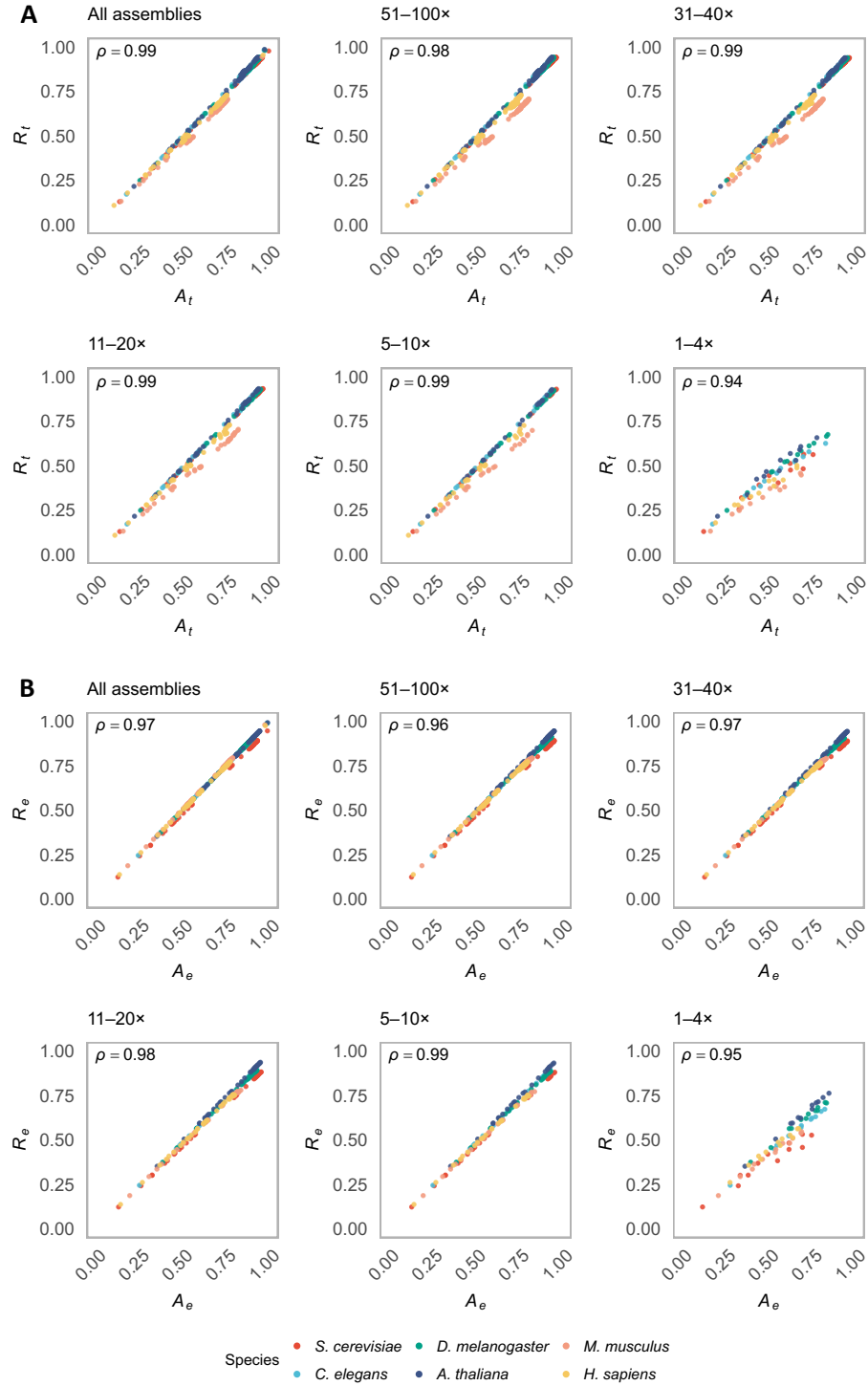


Figure 35. Correlation between relative and annotation-based transcript (A) and exon (B) scores in assemblies with decreasing coverage, as indicated in panel titles. R_t = relative transcript score, A_t = annotation-based transcript score, R_e = relative exon score, A_e = annotation-based exon score. Reproduced from Bodulić & Vlahoviček, 2026.

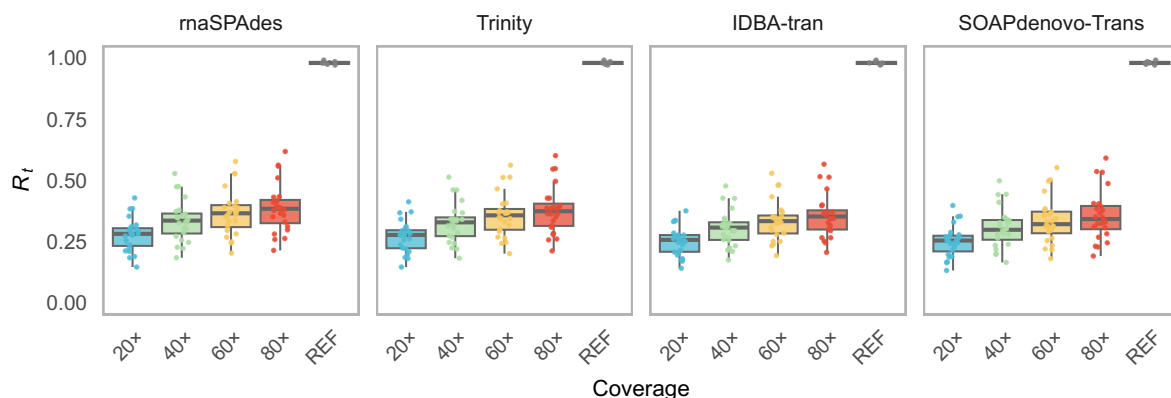


Figure 36. Distribution of relative CATS-rb transcript scores across assemblers and coverage in realistically simulated assemblies. R_t = relative transcript score, REF = reference transcriptome. Reproduced from Bodulić & Vlahoviček, 2026.

Next, CATS-rb was applied to assemblies from individual realistically simulated libraries alongside the corresponding reference transcriptome. In this setting, assemblies also displayed a strong correlation between relative and annotation-based scores (transcript: $\rho = 0.99$, 95% CI 0.98–0.99, exon: $\rho = 0.99$, 95% CI 0.98–0.99) (Figure 37).

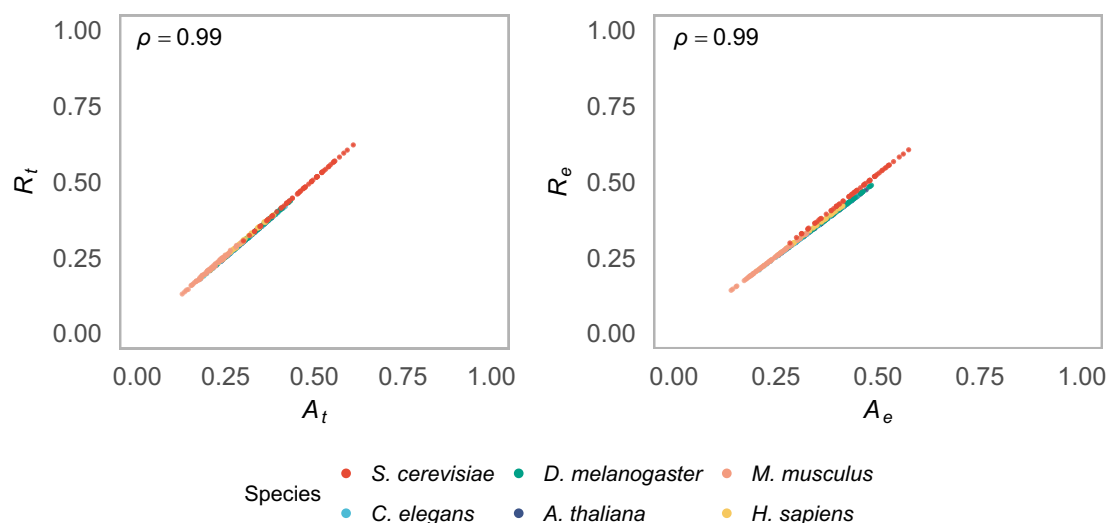


Figure 37. Correlation between relative and annotation-based scores in realistically simulated assemblies. R_t = relative transcript score, A_t = annotation-based transcript score, R_e = relative exon score, A_e = annotation-based exon score. Reproduced from Bodulić & Vlahoviček, 2026.

CATS-rb was also used to analyse assemblies from each public library paired with the corresponding reference transcriptome. The public dataset exhibited similarly strong correlations for relative transcript ($\rho = 0.98$, 95% CI 0.97–0.98) and exon ($\rho = 0.93$, 95% CI 0.91–0.95) scores (Figure 38).

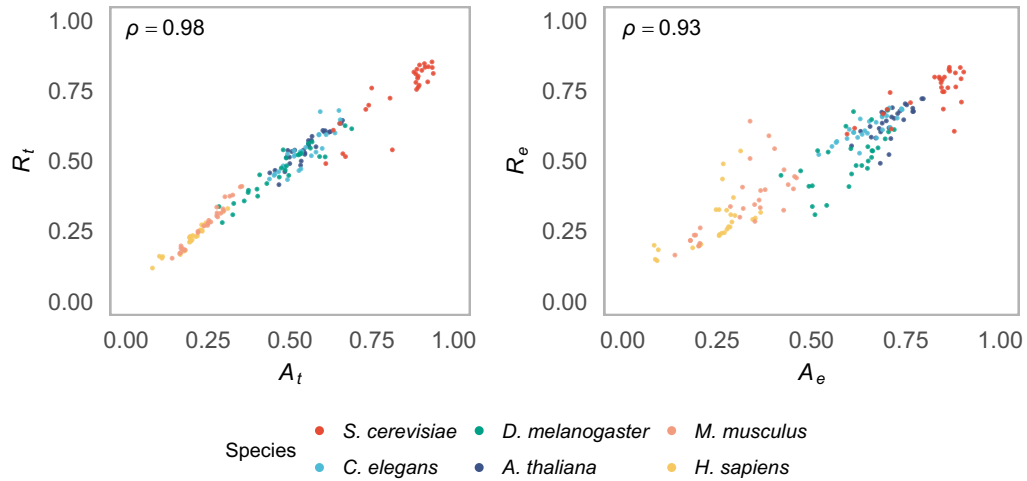


Figure 38. Correlation between relative and annotation-based scores in assemblies from public libraries. R_t = relative transcript score, A_t = annotation-based transcript score, R_e = relative exon score, A_e = annotation-based exon score. Reproduced from Bodulić & Vlahoviček, 2026.

Furthermore, CATS-rb was applied to native and mutated assemblies, the latter encompassing six levels of multiplicative mutations. CATS-rb successfully detected the introduced assembly errors, as reflected by a gradual decline in relative completeness scores with increasing mutation levels (Figure 39).

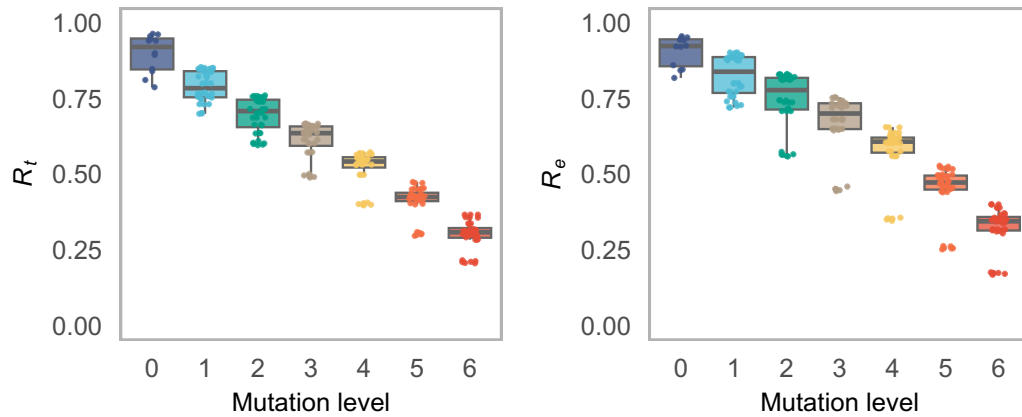


Figure 39. Distribution of CATS-rb relative scores in assemblies with multiplicative mutations. Mutation level 0 corresponds to native assemblies. R_t = relative transcript score R_e = relative exon score. Reproduced from Bodulić & Vlahoviček, 2026.

The consistency of assembly rankings was also assessed using measures of assembly completeness and accuracy within the controlled dataset. A strong association was observed between CATS-rb relative transcript scores and CATS-rf assembly scores in the full dataset (transcript: $\rho = 0.92$, 95% CI 0.91–0.93, exon: $\rho = 0.83$, 95% CI 0.81–0.85) and across all

assembly subsets, including the lowest-coverage subset (transcript: $\rho = 0.69$, 95% CI 0.54–0.79, exon: $\rho = 0.71$, 95% CI 0.59–0.81) (Figure 40).

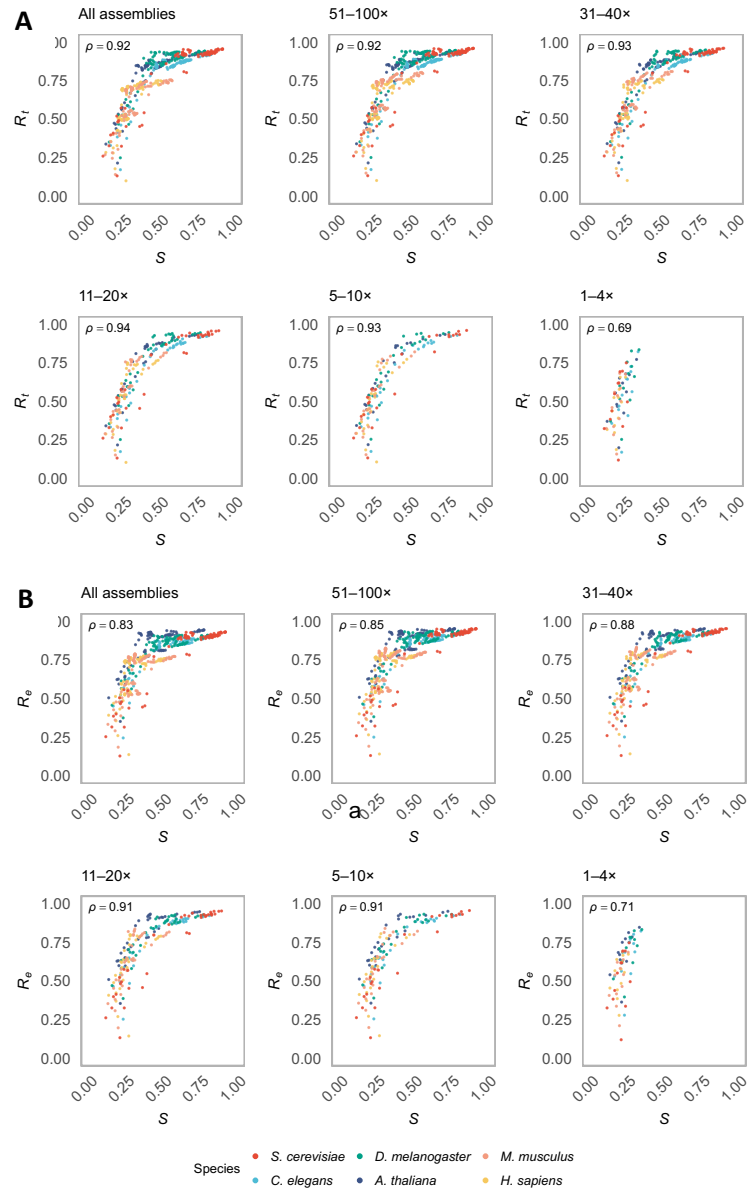


Figure 40. Correlation between CATS-rb relative transcript (A) and exon (B) scores with CATS-rf assembly scores across assemblies with decreasing coverage, as indicated in panel titles. R_t = relative transcript score, R_e = relative exon score, S = CATS-rf assembly score. Reproduced from Bodulić & Vlahoviček, 2026.

Finally, the performance of CATS-rb in detecting chimerism was assessed by artificially merging transcripts from reference transcriptome assemblies. Strong performance was demonstrated in classifying chimerism as a structural inconsistency, with high sensitivity (93.5%–96.8%) and specificity (97.0%–99.8%) observed across all species (Figure 41).

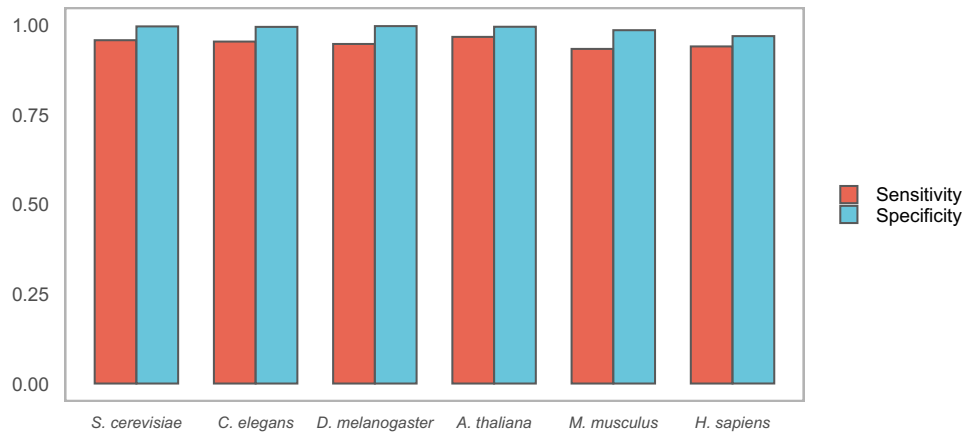


Figure 41. Performance of CATS-rb in classifying chimeric transcripts as structurally inconsistent.

4.2. *Sponge Transcriptome Assembly Evaluation*

Next, CATS was applied to evaluate transcriptomes generated from 259 stranded RNA-seq libraries representing 14 sponge species. Initial transcriptome reconstruction was performed de novo using six rnaSPAdes parameter configurations and Trinity with default settings. In parallel, genome-guided transcriptome assemblies were generated using Trinity. For each species, the best-performing rnaSPAdes assembly, the Trinity de novo assembly, and the Trinity genome-guided assembly were subsequently merged and subjected to redundancy reduction. The merged assemblies were then evaluated alongside the original assemblies to identify the optimal transcriptome for downstream lncRNA analyses.

4.2.1. *De Novo Transcriptome Assembly Quality*

All rnaSPAdes assemblies exhibited broadly similar quality metrics, whereas the Trinity assembly showed substantial differences in both accuracy and completeness (Figure 42). Overall, Trinity assemblies exhibited lower N_{50} values, integrity scores, and CATS-rf assembly scores. In contrast, Trinity consistently achieved significantly higher completeness metrics than any of the evaluated rnaSPAdes parameter settings.

The best rnaSPAdes assembly for each species was determined by calculating a z-score for selected quality metrics and averaging these scores across all metrics. Assembly performance was found to be species-dependent (Figure 43). Interestingly, the default rnaSPAdes assembly was not the best-performing option for any species.

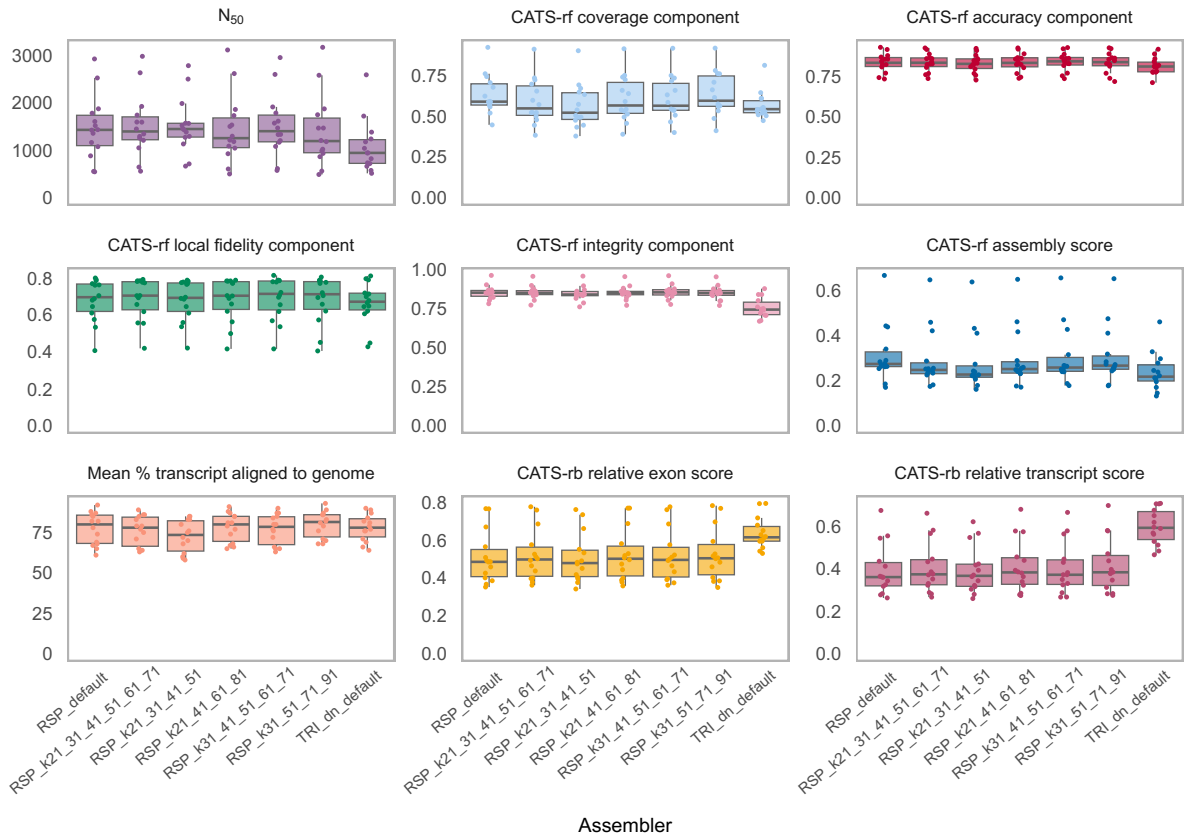


Figure 42. Distribution of selected quality metrics in de novo sponge transcriptome assemblies. RSP = rnaSPAdes, TRI_dn = Trinity de novo.

-0.16	-0.09	-0.09	-0.04	-0.05	-0.11	-0.08	-0.01	0.03	0.11	-0.17	0.07	-0.17	-0.15	RSP_default
0.13	0.14	-0.09	-0.08	0.07	0.05	-0.02	-0.19	0.09	0.06	0.02	0.20	-0.20	0.05	RSP_k21_31_41_51_61_71
-0.14	-0.08	-0.32	-0.53	-0.25	-0.41	-0.14	-0.56	-0.67	-0.76	-0.35	-0.52	-0.10	0.09	RSP_k21_31_41_51
0.03	-0.01	-0.07	0.14	0.14	0.08	-0.12	-0.02	0.01	0.14	0.06	0.26	-0.28	-0.07	RSP_k21_41_61_81
0.14	0.15	0.03	0.03	0.10	0.11	0.03	-0.09	0.10	0.23	-0.01	0.23	-0.06	-0.07	RSP_k31_41_51_61_71
-0.24	-0.15	0.19	0.26	0.16	0.11	-0.01	0.16	0.20	0.42	0.18	0.31	-0.16	0.06	RSP_k31_51_71_91
0.26	0.03	0.36	0.22	-0.18	0.17	0.34	0.71	0.23	-0.19	0.27	-0.54	0.97	0.10	TRI_dn_default
A. queenslandica	A. beatrix	A. aerophoba	C. candelabrum	C. crambe	D. avara	E. muelleri	H. caerulea	O. lobularis	P. ficiformis	S. officinalis	S. lacustris	S. ciliatum	T. wilhelma	

Figure 43. Distribution of z-scores derived from quality metrics in de novo sponge transcriptome assemblies. RSP = rnaSPAdes, TRI_dn = Trinity de novo.

4.2.2. Genome-guided and Combined Transcriptome Assembly Quality

Next, the best-performing rnaSPAdes assembly for each species was clustered with Trinity de novo and genome-guided assemblies using CD-HIT-EST with four parameter settings. All assemblies were subsequently compared using N_{50} and CATS-rb metrics. The total number of transcripts in clustered assemblies was generally comparable to that of individual de novo assemblies. Overall, combined assemblies exhibited the highest completeness scores, with peak completeness observed under the most conservative clustering parameters (Figure 44). Although genome-guided assemblies showed slightly lower completeness, they consistently achieved the highest contiguity and greatest fraction of transcript sequence aligned to the genome.

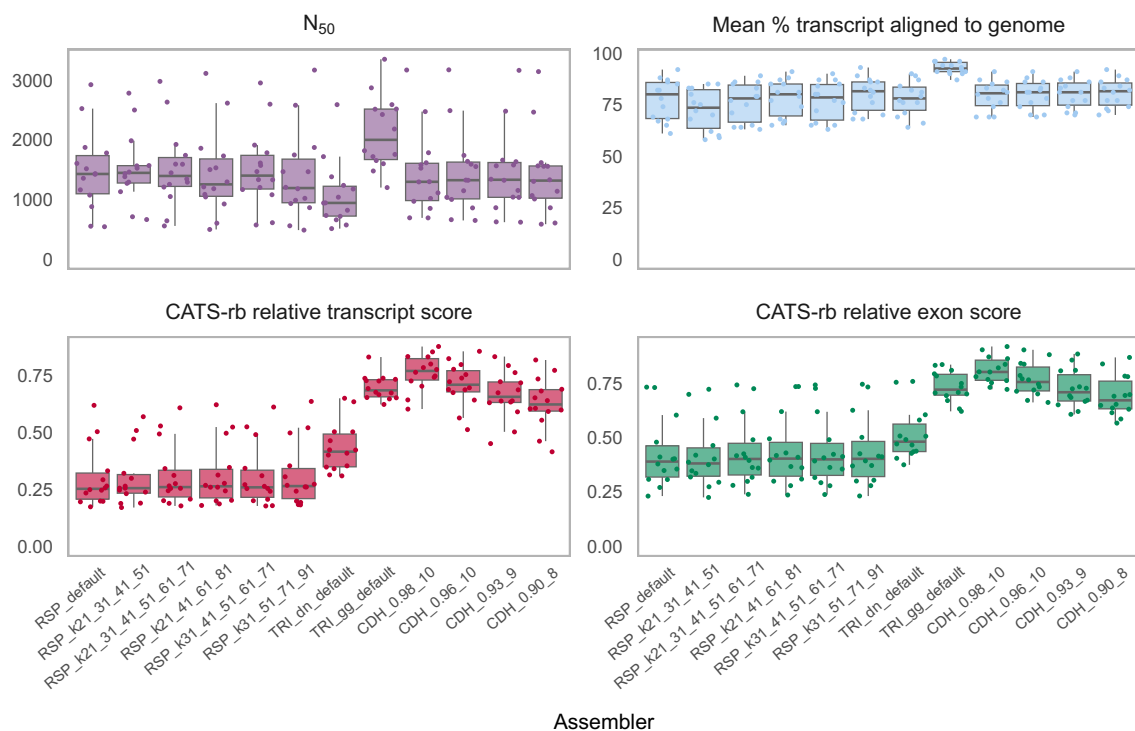


Figure 44. Distribution of selected quality metrics in all analysed sponge transcriptome assemblies. RSP = rnaSPAdes, TRI = Trinity, CDH = CD-HIT-EST.

Across all species, genome-guided and clustered transcriptome assemblies consistently outperformed all de novo assemblies (Figure 45). The optimal assembly for each species was selected using a strategy that prioritized assembly completeness while also accounting for redundancy. First, CATS-rb relative transcript scores were compared across all assemblies. The highest-scoring assembly was selected unless a less redundant assembly (defined by more stringent clustering parameters) or a genome-guided assembly demonstrated comparable completeness.

Using this approach, Trinity genome-guided assemblies were selected for five species: *A. queenslandica*, *A. beatrix*, *D. avara*, *E. muelleri*, and *S. officinalis*. The first clustering level (-c

0.98 -n 10) was chosen for seven species: *A. aerophoba*, *C. candelabrum*, *H. caerulea*, *P. ficiformis*, *S. ciliatum*, *S. lacustris*, and *T. wilhelma*. The second clustering level (-c 0.96 -n 10) was selected for *C. crambe* and *O. lobularis*.

0.27	0.20	0.26	0.62	0.25	0.29	0.17	0.20	0.50	0.47	0.25	0.33	0.23	0.20	RSP_default
0.27	0.19	0.26	0.57	0.24	0.30	0.17	0.23	0.51	0.47	0.25	0.32	0.26	0.19	RSP_k21_31_41_51
0.28	0.20	0.27	0.61	0.26	0.31	0.18	0.21	0.53	0.49	0.26	0.34	0.24	0.19	RSP_k21_31_41_51_61_71
0.28	0.20	0.27	0.62	0.26	0.31	0.18	0.20	0.52	0.50	0.26	0.35	0.24	0.19	RSP_k21_41_61_81
0.27	0.20	0.26	0.61	0.26	0.31	0.18	0.21	0.53	0.49	0.25	0.34	0.24	0.18	RSP_k31_41_51_61_71
0.27	0.19	0.27	0.64	0.26	0.31	0.18	0.20	0.52	0.50	0.26	0.35	0.24	0.18	RSP_k31_51_71_91
0.41	0.34	0.42	0.65	0.50	0.35	0.42	0.46	0.64	0.50	0.31	0.40	0.31	0.35	TRI_dn_default
0.73	0.83	0.65	0.74	0.73	0.67	0.73	0.68	0.69	0.65	0.65	0.62	0.67	0.74	TRI_gg_default
0.76	0.83	0.74	0.85	0.83	0.70	0.75	0.78	0.88	0.79	0.64	0.73	0.60	0.80	CDH_c0.98_n10
0.70	0.78	0.69	0.82	0.80	0.64	0.69	0.72	0.86	0.75	0.56	0.68	0.51	0.74	CDH_c0.96_n10
0.63	0.72	0.64	0.79	0.76	0.59	0.64	0.67	0.83	0.72	0.50	0.63	0.45	0.69	CDH_c0.93_n9
0.59	0.68	0.61	0.77	0.73	0.55	0.60	0.64	0.82	0.69	0.46	0.60	0.41	0.65	CDH_c0.90_n8
<i>A. queenslandica</i>	<i>A. beatrice</i>	<i>A. aerophoba</i>	<i>C. candelabrum</i>	<i>C. crambe</i>	<i>D. avara</i>	<i>E. muelleri</i>	<i>H. caerulea</i>	<i>O. lobularis</i>	<i>P. ficiformis</i>	<i>S. officinalis</i>	<i>S. lacustris</i>	<i>S. ciliatum</i>	<i>T. wilhelma</i>	

Figure 45. Distribution of CATS-rb relative transcript scores in de novo sponge transcriptome assemblies. RSP = rnaSPAdes, TRI = Trinity, CDH = CD-HIT-EST.

4.3. Genome Assembly Decontamination

Considering the holobiont nature of sponges, a comprehensive decontamination workflow was required for sponge genome assemblies prior to lncRNA analysis. To address this, a genome decontamination pipeline specifically designed for non-model organisms was developed and applied to all analysed sponge genomes.

4.3.1. Genome Decontamination Pipeline Validation

The developed decontamination pipeline combines complementary lines of evidence to improve contaminant classification accuracy. Specifically, this approach integrates a compositional score that evaluates k-mer profiles and GC content, a protein score based on taxonomic classification of encoded proteins, and a nucleotide score derived from k-mer taxonomic hits.

The genome decontamination pipeline was validated using a spike-in dataset encompassing bona fide sponge sequences alongside a diverse set of representative contaminants. The pipeline displayed high accuracy in classifying contaminant scaffolds (96.8%), with strong precision (99.6%) and recall (90.8%) (Figure 46). Among the individual components, the compositional score exhibited the poorest performance (accuracy 74.6%, precision 59.5%, recall 77.1%), while protein and nucleotide taxonomy scores displayed significantly stronger metrics (protein: accuracy 96.2%, precision 98.2%, recall 90.5%; nucleotide: accuracy 96.0%,

precision 99.4%, recall 88.8%). The integrated score achieved the highest accuracy, outperforming each individual component, although performance gains relative to the protein and nucleotide scores were modest. All sponge mitochondrial sequences were accurately classified despite their prokaryotic characteristics. Furthermore, one (6.7%) bacterial plasmid sequence was incorrectly classified as non-contaminant.

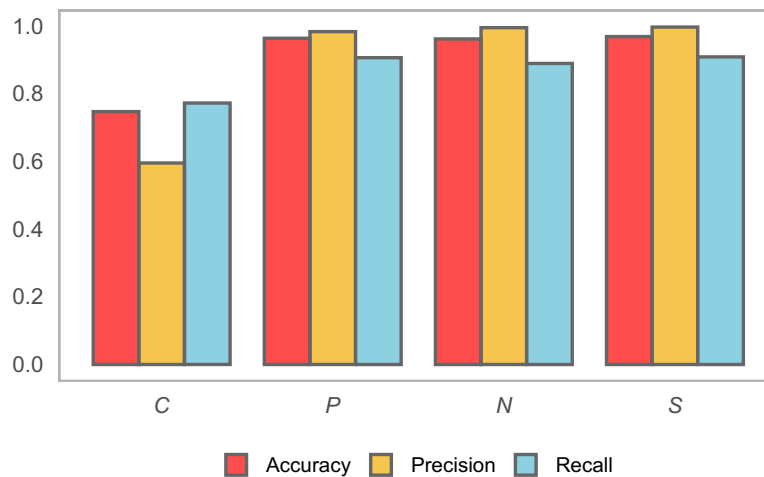


Figure 46. Accuracy, precision, and recall for individual scoring components and for the integrated contamination score within the genome decontamination pipeline. *C* = compositional score, *P* = protein taxonomy score, *N* = nucleotide taxonomy score, *S* = contamination score. Reproduced from Bodulić & Vlahoviček, 2026.

The intersection analysis suggested strong agreement among the scoring components (Figure 47). Most non-contaminant sequences were correctly classified by all three components (6649, 72.4% of true non-contaminants). Similarly, 3120 sequences (67.1% of true contaminants) were accurately classified as contaminants by all three scores, while 891 sequences (19.2% of true contaminants) were correctly classified as contaminants by protein and nucleotide taxonomic scores, but not by the compositional score. Among scaffolds supported by a single contamination signal, the compositional score incorrectly classified 2430 sequences as contaminants (26.5% of true non-contaminants), indicating a higher false positive rate when used alone. However, 257 contaminant sequences (5.5% of true contaminants) were identified by the compositional score but incorrectly determined as non-contaminants by both protein and nucleotide scores. Generally, the length distribution across the analysed groups did not imply a strong length-associated bias. An exception was observed for contaminant sequences missed by the nucleotide score but detected by both compositional and protein components, which were notably short (median 1.79 kb, IQR 3.02 kb).

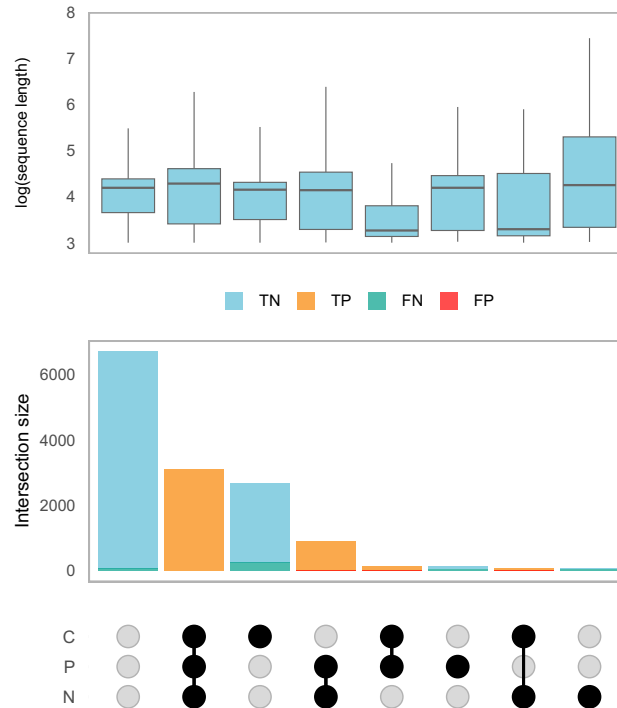


Figure 47. UpSet plot showing agreement among scoring components within the genome decontamination pipeline. Bar heights indicate the number of sequences in each intersection, while boxplots summarise sequence length distributions for each intersection. TN = true negatives, TP = true positives, FN = false negatives, FP = false positives, C = compositional score, P = protein taxonomy score, N = nucleotide taxonomy score. Reproduced from Bodulić & Vlahoviček, 2026.

Recall varied across taxonomic groups and scoring components, with moderate differences observed among taxonomic domains (Figure 48). The integrated score displayed consistently high recall across most phyla. This pattern was also observed for the protein and nucleotide taxonomy scores. In contrast, the compositional score exhibited lower recall in several prokaryotic and eukaryotic phyla. Notably, the compositional score successfully identified contaminants from the Filasterea phylum that were not captured by the taxonomy-based approaches. Viral phyla showed low recall across all score components and the integrated score.

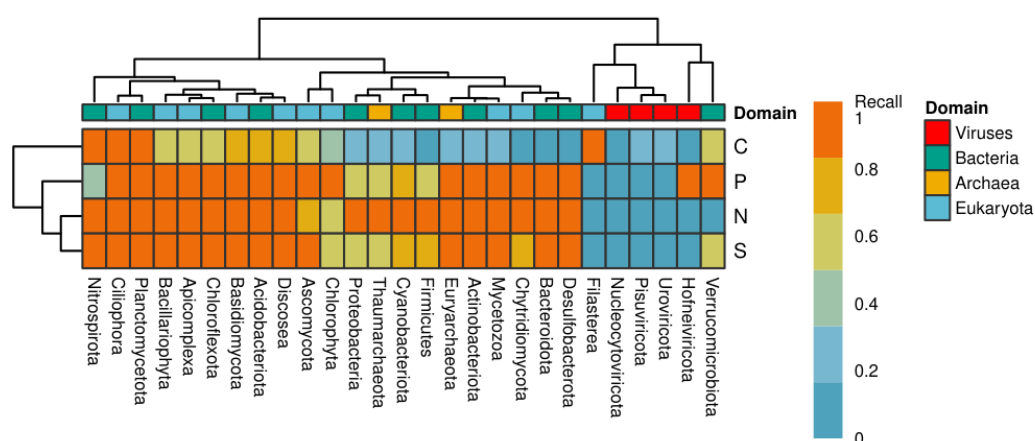


Figure 48. Genome decontamination pipeline recall across contaminant phyla for individual scoring components and the final integrated score, with hierarchical clustering applied to both phyla and scoring methods. Reproduced from Bodulić & Vlahoviček, 2026.

Taxonomic assignment accuracy for contaminant sequences was also evaluated. An assignment was considered accurate when the designated lineage matched any taxonomic rank within the true lineage of the corresponding taxon. Based on this criterion, 4223 contaminant sequences (97.7%) were classified correctly, with 4068 sequences (96.3%) accurately resolved to the genus level. Only 63 sequences (1.5%) remained unclassified and were assigned as unknown. Strong classification performance was observed across the vast majority of contaminant taxa, with low accuracy restricted to only two phyla: Filasterea (0%) and Thaumarchaeota (16.7%) (Figure 49). All misclassified Thaumarchaeota sequences were assigned to alternative archaeal genera, whereas Filasterea sequences were primarily misclassified as fungal or algal taxa.

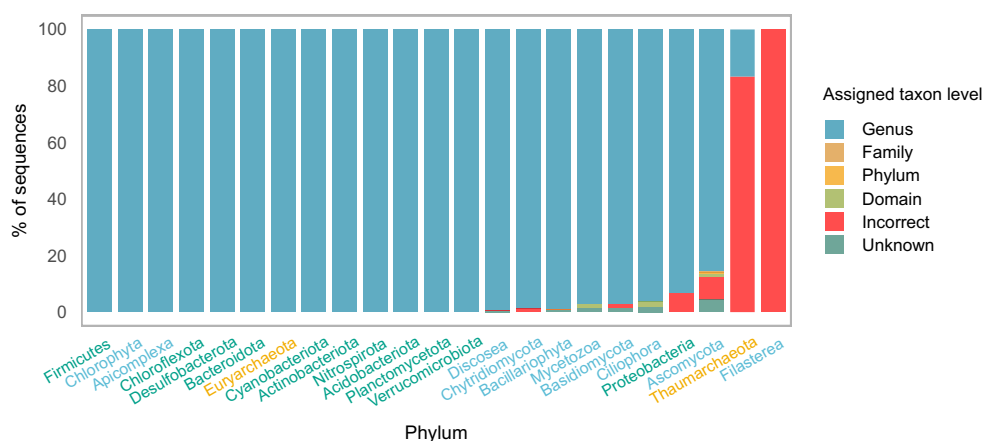


Figure 49. Accuracy of taxonomic assignment for contaminant sequences, showing the proportion of sequences resolved to different taxonomic ranks as well as incorrect and unknown classifications. X-axis labels represent contaminant phyla and are colored by taxonomic domains analogous to Figure 48. Reproduced from Bodulić & Vlahoviček, 2026.

4.3.2. Decontamination of the Analysed Sponge Genome Assemblies

Next, the genome decontamination pipeline was applied to the analysed 14 sponge genome assemblies. The analysis revealed varying levels of contamination among datasets, as indicated by the number of scaffolds identified as contaminants (Table 13). The mean proportion of contaminant scaffolds was 8.3% (SD 10.3%), with *A. beatrix*, *C. candelabrum*, and *O. lobularis* showing no detectable contamination. The highest proportions of contaminant scaffolds were observed in *S. officinalis* (40, 37.8%), *C. crambe* (15, 12.1%), and *S. ciliatum* (48, 7.8%). When contamination was assessed based on the proportion of contaminated genome length, substantially lower levels were observed overall (mean 0.2%, SD 0.3%), with the highest level detected in *A. queenslandica* (1.22%). Mean contaminant length was 11.1 kb (IQR 20.8 kb), with the longest contaminants found in *H. caerulea* ($N = 2$, 206.4 and 179.2 kb). The number of complete Metazoan BUSCOs did not change significantly following decontamination in any analysed species.

Table 13. Summary statistics of contaminant sequences identified across the 14 analysed sponge genomes.

Species	Contaminant scaffolds			N complete BUSCOs (Metazoa)	
	N (%)	Length (Mb, %)	Length median (IQR), kb	Original genome	Decontaminated genome
<i>A. queenslandica</i>	191 (4.9%)	2.05 (1.22%)	6.6 (10.4)	529	529
<i>A. beatrix</i>	0 (0.0%)	0.00 (0.0%)	/	483	483
<i>A. aerophoba</i>	7 (4.4%)	0.27 (0.17%)	39.5 (28.3)	526	526
<i>C. candelabrum</i>	0 (0.0%)	0.00 (0.00%)	/	590	590
<i>C. crambe</i>	15 (12.1%)	0.24 (0.17%)	17.5 (8.4)	569	568
<i>D. avara</i>	8 (7.4%)	0.13 (0.02%)	9.1 (14.3)	548	548
<i>E. muelleri</i>	8 (4.8%)	0.19 (0.08%)	16.2 (27.1)	560	560
<i>H. caerulea</i>	2 (0.4%)	0.39 (0.18%)	192.8 (14.0)	519	518
<i>O. lobularis</i>	0 (0.0%)	0.00 (0.00%)	/	569	569
<i>P. ficiformis</i>	1 (2.8%)	0.01 (0.03%)	6.2 (0)	558	558
<i>S. officinalis</i>	40 (38.1%)	0.63 (0.14%)	12.8 (21.7)	556	556
<i>S. lacustris</i>	4 (5.7%)	0.05 (0.02%)	12.4 (9.4)	456	456
<i>S. ciliatum</i>	48 (7.8%)	1.27 (0.24%)	23.2 (23.2)	514	513
<i>T. wilhelma</i>	13 (2.3%)	0.35 (0.27%)	3.0 (18.1)	509	509

Mb = megabase, kb = kilobase

Next, taxonomic classification was performed on the detected contaminant scaffolds. The contaminant profiles were predominantly composed of bacterial phyla (Figure 50). Among these, Pseudomonadota was identified as the most abundant phylum, with 186 scaffolds detected across eight sponge species. Furthermore, two archaeal phyla were identified. This included Nitrososphaerota and Methanobacteriota, each with two scaffolds found in two species. Among eukaryotic contaminants, several algal taxa were detected, primarily belonging to green algae (Streptophyta, 11 scaffolds detected in 6 species). The greatest contaminant diversity was observed in *S. ciliatum*, which harbored 10 distinct phyla, including Pseudomonata (13 scaffolds) and Bacteroidota (10 scaffolds). High contaminant diversity was also detected in *A. queenslandica*, with eight identified contaminant phyla, five of which were

bacterial. Similarly, elevated contaminant diversity was observed in *S. officinalis* and *C. crambe*, each of which harbored six contaminant phyla.

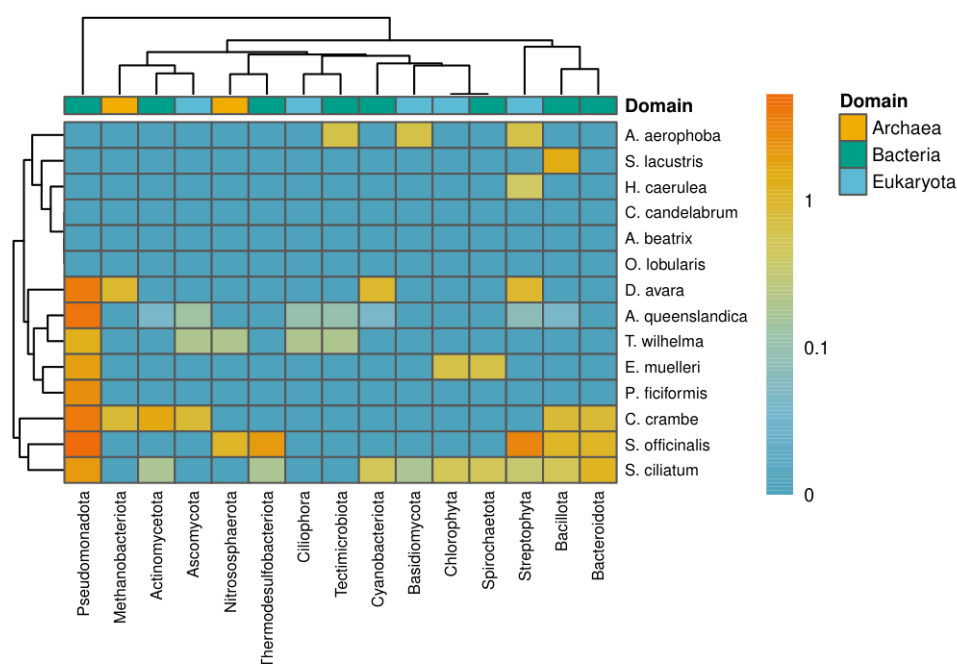


Figure 50. Proportion of contaminant scaffolds across sponge genome assemblies, with hierarchical clustering applied to contaminant phyla and sponge species. Adapted from Bodulić & Vlahoviček, 2026.

Next, concordance among the three contamination-scoring components was assessed for each sponge assembly. In contrast to the validation dataset, the majority of contaminant scaffolds identified across all sponge assemblies were supported by two of the three scoring components (Figure 51). Only 39 scaffolds (11.6%) were classified as contaminants by all three scores. The largest proportion of contaminant scaffolds (301, 89.3%) was identified exclusively by the compositional and protein scores, whereas 72 scaffolds (22.2%) were supported by compositional and nucleotide scores. Consistent with observations from the validation dataset, scaffolds were most frequently flagged as contaminants by the compositional score.

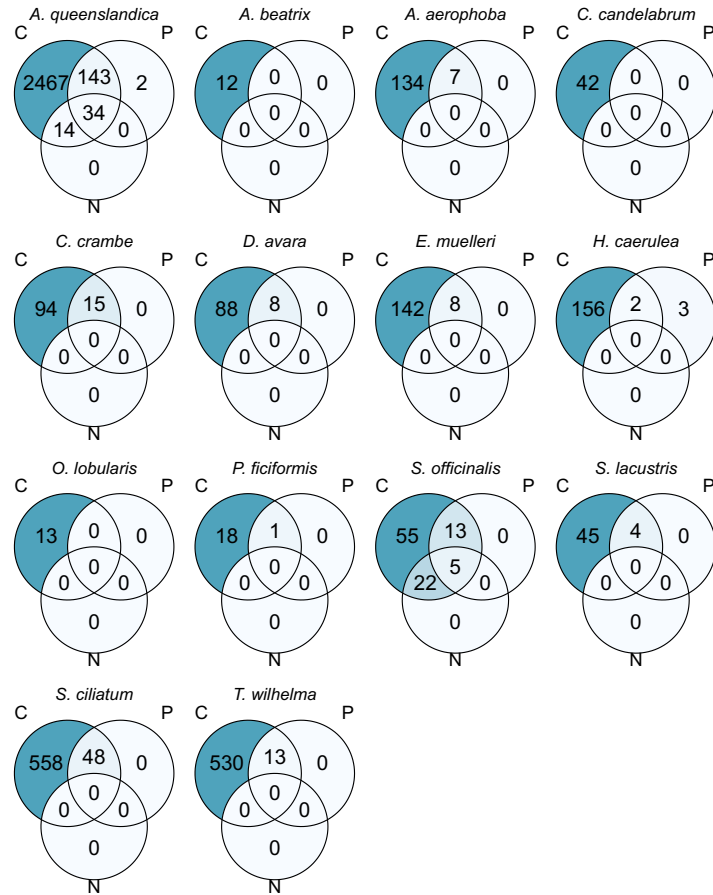


Figure 51. Intersection analysis illustrating agreement among the compositional, protein, and nucleotide scoring components for each genome assembly. C: compositional score, P: protein taxonomy score, N: nucleotide taxonomy score. Adapted from Bodulić & Vlahoviček, 2026.

4.4. Sponge Genome Annotation

Next, comprehensive genome annotation was performed across the analysed sponge genomes, including detailed characterization of TEs and structural and functional annotation of PCGs.

4.4.1. Repeat Annotation

Repeat annotation was conducted using a comprehensive pipeline designed to detect a broad range of TEs. Considerable variation was observed in the proportion of genome length occupied by repetitive elements among the analysed sponge genomes (Table 14). The proportion of repetitive sequences ranged from 20.3% in *O. lobularis* to 72.5% in *S. officinalis*. High proportions of repetitive DNA were also detected in *D. avara* (70.9%), *S. ciliatum* (70.7%), and *H. caerulea* (69.0%), whereas comparatively lower repeat content was observed in *T. wilhelma* (38.9%), *C. crambe* (44.3%), and *A. queenslandica* (44.5%).

Across all analysed genomes, DNA transposons represented the dominant repeat class and accounted for the largest proportion of genome length, ranging from 10.8% in *O. lobularis* to

44.9% in *D. avara*. Similarly high proportions of DNA transposons were identified in *S. officinalis* (44.1%), *H. caerulea* (42.0%), *S. lacustris* (41.9%), and *P. ficiformis* (40.0%). Furthermore, LTRs generally constituted the second most abundant repeat category, occupying up to 23.9% of genome length in both *S. officinalis* and *S. ciliatum*, followed by *D. avara* (20.5%) and *H. caerulea* (18.7%). LINE elements displayed substantial variation among species, contributing between 0.7% (*O. lobularis*) and 16.2% (*S. ciliatum*) of total genome length. Elevated proportions of LINEs were also observed in *C. candelabrum* (9.4%), *A. aerophoba* (9.2%), and *H. caerulea* (9.1%). In contrast, SINE elements consistently occupied only a minor fraction of the genome, with the highest proportion detected in *S. ciliatum* (3.3%). Helitron elements were identified at very low frequencies across all assemblies and contributed negligibly to overall genome length.

Table 14. Relative proportions (%) of major repeat classes identified in the analysed sponge genomes.

Species	Total	DNA	LTR	LINE	SINE	Helitron
<i>A. queenslandica</i>	44.5	28.8	10.9	2.2	0.6	0.0
<i>A. beatrix</i>	52.0	39.8	9.2	1.2	0.2	0.0
<i>A. aerophoba</i>	48.6	29.9	10.0	9.2	0.4	0.0
<i>C. candelabrum</i>	53.8	29.4	14.4	9.4	0.3	0.0
<i>C. crambe</i>	44.3	24.9	13.4	2.7	1.6	0.0
<i>D. avara</i>	70.9	44.9	20.5	8.6	0.8	0.0
<i>E. muelleri</i>	58.0	38.1	14.4	6.3	0.5	0.2
<i>H. caerulea</i>	69.0	42.0	18.7	9.1	0.6	0.1
<i>O. lobularis</i>	20.3	10.8	5.9	0.7	0.0	0.0
<i>P. ficiformis</i>	58.5	40.0	12.4	6.3	0.7	0.0
<i>S. officinalis</i>	72.5	44.1	23.9	6.7	0.2	0.0
<i>S. lacustris</i>	65.3	41.9	16.9	7.7	0.2	0.0
<i>S. ciliatum</i>	70.7	29.4	23.9	16.2	3.3	0.0
<i>T. wilhelma</i>	38.9	22.7	10.7	3.7	0.8	0.1

LTR = Long terminal repeat retrotransposons, LINE = long interspersed nuclear elements, SINE = short interspersed nuclear elements

Repeat composition was further assessed by analysing the genomic abundance of major repeat classes in each genome assembly (Table 15). Consistent with the previous analysis, substantial variation in repeat abundance was observed among species. Across all analysed species, DNA transposons represented the dominant repeat category and accounted for the largest proportion of annotated repeats, comprising between 34.0% (*S. ciliatum*) and 53.2% (*A. queenslandica*) of all repeat annotations. LTR retrotransposons were generally the second most abundant repeat class, contributing up to 18.4% of all repeats in *O. lobularis* and 17.5% in *C. crambe*. Elevated proportions of LTR elements were also detected in *S. ciliatum* (17.0%) and *T. wilhelma* (16.2%). Furthermore, LINE elements showed marked variability among species and accounted for between 1.3% (*A. beatrix*) and 10.3% (*S. ciliatum*) of all annotated repeats. Particularly high proportions of LINEs were additionally identified in *C. candelabrum* (9.5%), *A. aerophoba* (8.4%), and *H. caerulea* (6.9%). SINE elements were comparatively rare in all assemblies, with the highest proportion observed in *S. ciliatum* (3.7%). Helitron elements were consistently detected at extremely low frequencies and represented less than 0.1% of

repeat annotations in nearly all species. A substantial proportion of repeat elements across all genomes remained unclassified, accounting for between 25.6% (*O. lobularis*) and 36.0% (*D. avara*) of all annotated repeats.

Table 15. Distribution of major repeat classes in the analysed sponge genome assemblies.

Species	Total	DNA	LTR	LINE	SINE	Helitron	Unknown
<i>A. queenslandica</i>	531895	282893 (53.2%)	62791 (11.8%)	20747 (3.9%)	6662 (1.3%)	22 (0.0%)	158780 (29.9%)
<i>A. beatrix</i>	379428	200222 (52.8%)	37678 (9.9%)	4759 (1.3%)	1268 (0.3%)	2 (0.0%)	135499 (35.7%)
<i>A. aerophoba</i>	626124	303236 (48.4%)	79150 (12.6%)	52373 (8.4%)	4164 (0.7%)	81 (0.0%)	187120 (29.9%)
<i>C. candelabrum</i>	700635	291938 (41.7%)	108215 (15.4%)	66443 (9.5%)	4139 (0.6%)	73 (0.0%)	229827 (32.8%)
<i>C. crambe</i>	457765	189823 (41.5%)	80181 (17.5%)	20295 (4.4%)	9786 (2.1%)	25 (0.0%)	157655 (34.4%)
<i>D. avara</i>	2947566	1254222 (42.6%)	432700 (14.7%)	177716 (6.0%)	21378 (0.7%)	49 (0.0%)	1061501 (36.0%)
<i>E. muelleri</i>	868494	425697 (49.0%)	112091 (12.9%)	50865 (5.9%)	5664 (0.7%)	872 (0.1%)	273305 (31.5%)
<i>H. caerulea</i>	1148172	529445 (46.1%)	148461 (12.9%)	79411 (6.9%)	7093 (0.6%)	311 (0.0%)	383451 (33.4%)
<i>O. lobularis</i>	55851	29433 (52.7%)	10279 (18.4%)	1610 (2.9%)	252 (0.5%)	0 (0.0%)	14277 (25.6%)
<i>P. ficiformis</i>	1137936	577560 (50.8%)	112546 (9.9%)	65835 (5.8%)	8673 (0.8%)	25 (0.0%)	373297 (32.8%)
<i>S. officinalis</i>	2827348	1280679 (45.3%)	389106 (13.8%)	144526 (5.1%)	5838 (0.2%)	69 (0.0%)	1007130 (35.6%)
<i>S. lacustris</i>	1022669	503309 (49.2%)	130820 (12.8%)	63469 (6.2%)	3732 (0.4%)	428 (0.0%)	320911 (31.4%)
<i>S. ciliatum</i>	2732873	928251 (34.0%)	465665 (17.0%)	280530 (10.3%)	100680 (3.7%)	153 (0.0%)	957594 (35.0%)
<i>T. wilhelma</i>	412851	187729 (45.5%)	66729 (16.2%)	21136 (5.1%)	996 (0.2%)	233 (0.1%)	136028 (32.9%)

LTR = Long terminal repeat retrotransposons, LINE = long interspersed nuclear elements, SINE = short interspersed nuclear elements

4.4.2. Protein-coding Gene Annotation

Protein-coding gene annotation was performed using evidence derived from the available stranded RNA-seq libraries together with homologous metazoan proteins. Considerable variation in gene architecture was observed across the analysed sponge genomes (Table 16). The number of predicted PCGs ranged from 14881 in *C. crambe* to 28699 in *E. muelleri*. Similarly, transcript counts varied substantially among species, with the lowest number again detected in *C. crambe* (15942) and the highest in *E. muelleri* (33138). Exon counts also differed markedly, ranging from 37853 in *A. beatrix* to 185536 in *S. ciliatum*. The proportion of genome length occupied by annotated genes ranged from 16.5% in *D. avara* to 59.4% in *S. ciliatum*. High genic proportions were also observed in *H. caerulea* (57.5%) and *O. lobularis* (52.0%), whereas comparatively lower values were identified in *S. officinalis* (23.3%) and *T. wilhelma* (27.2%). Exonic regions accounted for a smaller fraction of genome length across all species,

ranging from 4.5% in *D. avara* to 39.3% in *O. lobularis*. Intron content exhibited substantial variation and represented the largest component of genic sequence in most assemblies, contributing between 11.4% (*A. beatrix*) and 47.7% (*S. ciliatum*) of total genome length. Particularly high intronic proportions were additionally observed in *H. caerulea* (39.4%), *A. aerophoba* (32.3%), and *P. ficiformis* (29.5%).

Table 16. Structural annotation statistics and relative genomic proportions (%) of annotated PCG features across the analysed sponge genomes.

Species	N			% genome length		
	Genes	Transcript	Exons	Genes	Exons	Introns
<i>A. queenslandica</i>	24555	28010	135827	32.8	17.4	15.4
<i>A. beatrix</i>	17588	19395	37853	33.6	22.2	11.4
<i>A. aerophoba</i>	22597	26219	118075	50.1	17.8	32.3
<i>C. candelabrum</i>	22494	24610	91745	30.8	12.2	18.6
<i>C. crambe</i>	14881	15942	78244	30.4	15.3	15.1
<i>D. avara</i>	16953	17418	98950	16.5	4.5	12.0
<i>E. muelleri</i>	28699	33138	133897	33.7	13.7	20.0
<i>H. caerulea</i>	23588	26823	117174	57.5	18.1	39.4
<i>O. lobularis</i>	15713	18333	126920	52.0	39.3	12.7
<i>P. ficiformis</i>	20765	23464	126199	44.4	14.9	29.5
<i>S. officinalis</i>	23683	27128	143519	23.3	6.8	16.5
<i>S. lacustris</i>	26035	29787	120131	30.3	12.2	18.1
<i>S. ciliatum</i>	26678	31902	185536	59.4	11.7	47.7
<i>T. wilhelma</i>	15417	16680	59812	27.2	15.1	12.1

4.5. Long Non-coding RNA Analysis

The processed transcriptomic and genomic datasets were used to generate a comprehensive catalogue of lncRNAs, accompanied by detailed annotation analyses in 14 sponge genome assemblies.

4.5.1. Long Non-coding RNA Identification

Identification of lncRNAs was performed using a stringent multi-step filtering pipeline designed to retain high-confidence lncRNAs while excluding potential PCTs and assembly artefacts. The pipeline was applied consistently across all analysed sponge species and produced broadly similar filtering patterns (Figure 52).

In all species, the initial set of assembled transcripts was progressively reduced through sequential filtering steps, including removal of transcripts lacking structural integrity, short transcripts, low-expression transcripts, and sequences with coding potential or similarity to known protein elements. The largest reduction in transcript number generally occurred during the structural consistency and coding-potential filtering stages. In several species, transcript numbers decreased only modestly following mapping and structural consistency filtering, while other species exhibited pronounced reductions at these stages. Subsequent coding-potential filtering represented one of the most substantial reduction steps across most taxa. After this filtering step, only a small proportion of transcripts were excluded due to similarity to known proteins or overlap with protein-coding exons.

Species with larger transcriptomes, such as *A. queenslandica*, *H. caerulea*, and *D. avara*, retained comparatively high numbers of candidate lncRNAs after filtering, whereas species with smaller transcript sets, including *C. candelabrum* and *P. ficiformis*, showed lower final counts. Across all analysed species, the final high-confidence lncRNA repertoires ranged from 3761 transcripts in *P. ficiformis* to 37489 transcripts in *E. muelleri*, with most species retaining between 8000 and 32000 candidate lncRNAs after all filtering steps.

4.5.2. General Long Non-coding RNA Characteristics

General lncRNA characteristics were analysed and compared with protein-coding sequences across the analysed sponge genomes. The comparison included three transcript sets: identified lncRNAs, PCTs derived from the same transcriptome assemblies used for lncRNA identification, and PCTs defined by BRAKER3 genome annotation, which incorporated de novo prediction, protein homology, and RNA-seq evidence.

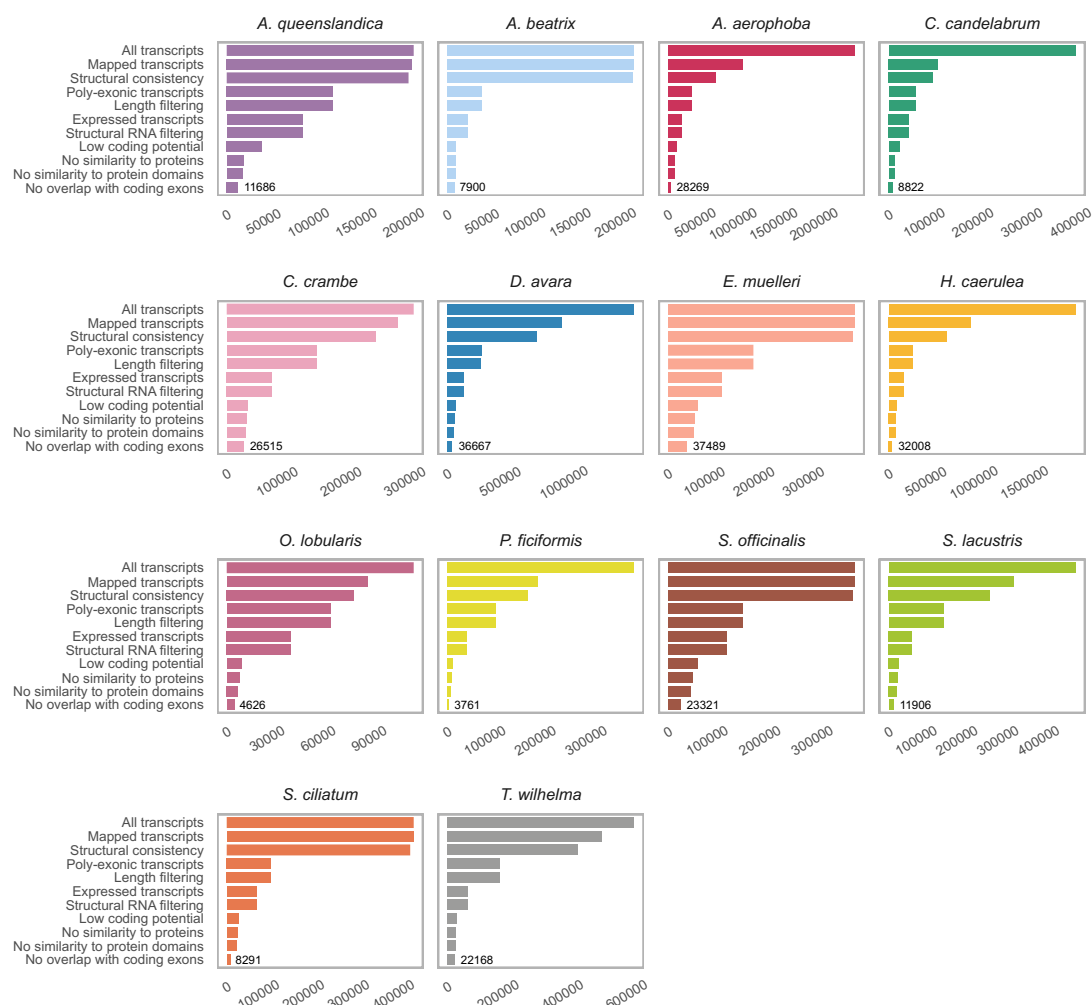


Figure 52, Sequential filtering pipeline used for identification of lncRNAs across the analysed sponge species. Bar plots show the number of transcripts retained after each filtering step.

The analysed transcript classes showed similar trends in length distributions across all sponge species (Figure 53). Notably, lncRNA transcripts were consistently shorter than both transcriptome-derived PCTs ($p < 0.001$) and BRAKER3-annotated PCTs ($p = 0.012$ in *A. beatrix*; $p < 0.001$ in all other species). Median lncRNA transcript lengths ranged from 320 bp in *C. candelabrum* to 644 bp in *A. beatrix*, whereas PCTs derived from transcriptome assemblies generally ranged from 641 bp in *S. ciliatum* to 3120 bp in *A. beatrix*. Assembly-derived PCTs were generally longer than BRAKER3-derived PCTs ($p < 0.001$). An exception was observed in *S. ciliatum*, which exhibited substantially longer BRAKER3 coding transcripts compared with transcriptome-derived mRNAs (median lengths 915 bp and 641 bp, $p < 0.001$). Identical trends were observed when analysing gene lengths, with lncRNA genes being consistently shorter than both assembly-derived PCGs ($p < 0.001$) and BRAKER3 PCGs ($p = 0.034$ for *A. beatrix*; $p < 0.001$ for all other species). Median lncRNA gene lengths ranged from 605 bp in *C. candelabrum* to 1414 bp in *A. beatrix*, whereas median lengths of assembly-derived PCGs ranged from 2088 bp in *A. queenslandica* to 7058 bp in *H. caerulea*.

In contrast to transcript and gene lengths, exon and intron lengths were generally more comparable among lncRNAs, assembly-derived PCTs, and BRAKER3 PCTs across all analysed sponge species. Median exon lengths showed relatively limited variation among transcript classes, typically ranging between 110 and 180 bp in most taxa. However, *A. beatrix* represented a notable exception, with BRAKER3 PCTs exhibiting significantly longer exons than both transcriptome-derived PCTs (medians 313 vs. 256 bp, $p < 0.001$) and lncRNAs (medians 313 vs. 239 bp, $p < 0.001$). Intron lengths likewise exhibited broadly similar distributions among transcript classes, although several species-specific patterns were observed. In *H. caerulea* and *S. ciliatum*, lncRNAs displayed significantly shorter introns than both assembly-derived and BRAKER3 PCTs ($p < 0.001$). On the other hand, lncRNA intron lengths (median 344.5 bp) in *S. ciliatum* were significantly lower when compared to assembly-derived (median 710 bp, $p < 0.001$) and BRAKER3 introns (median 719 bp, $p < 0.001$).

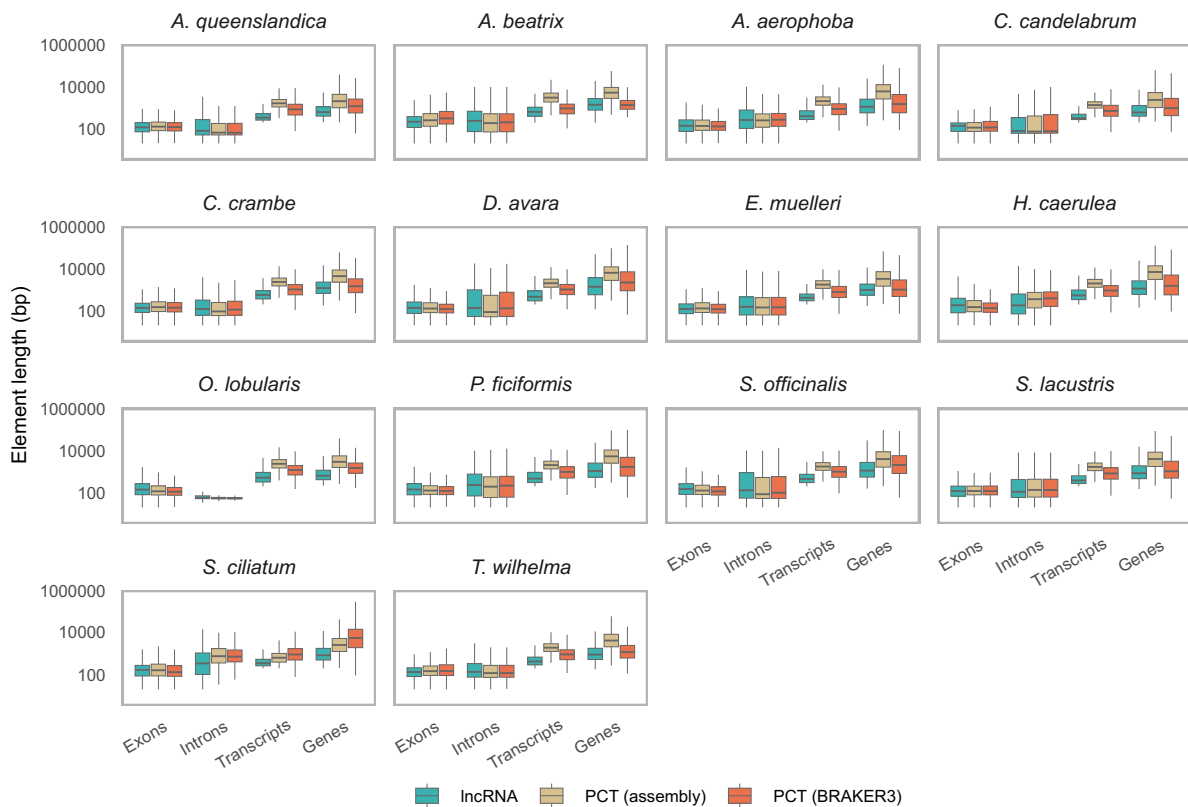


Figure 53. Distribution of exon, intron, transcript, and gene lengths among identified lncRNAs, assembly-derived PCTs, and BRAKER3-derived PCTs across the analysed sponge species. lncRNA = long non-coding RNA, PCT = protein-coding transcript.

The analysed transcript classes exhibited similar patterns in exon number per transcript across all sponge species (Figure 54). Notably, only poly-exonic transcripts were included in this analysis to ensure comparability among transcript classes, as mono-exonic lncRNAs were excluded during the lncRNA identification pipeline. In nearly all taxa, lncRNAs contained significantly fewer exons than both assembly-derived ($p < 0.001$) and BRAKER3 PCTs ($p < 0.001$). Median exon counts in lncRNAs typically ranged from two to three exons per

transcript, whereas assembly-derived PCTs exhibited median exon counts between four and nine. Exon number distributions were highly similar between assembly-derived and BRAKER3 PCTs in most species. Again, a notable exception was observed in *S. ciliatum*, where assembly-derived PCTs displayed fewer exons than BRAKER3-derived PCTs, although exon numbers remained substantially higher than those of lncRNAs (median values of 6 vs. 2 exons per transcript, $p < 0.001$). Although most lncRNAs exhibited relatively low exon numbers, substantial variation was still observed within species. Maximum exon counts in lncRNAs ranged from nine exons in *S. ciliatum* to 64 exons in *T. wilhelma*. However, these values remained consistently lower than those observed in PCTs, which frequently exceeded 100 exons in both assembly- and BRAKER3-derived PCTs.

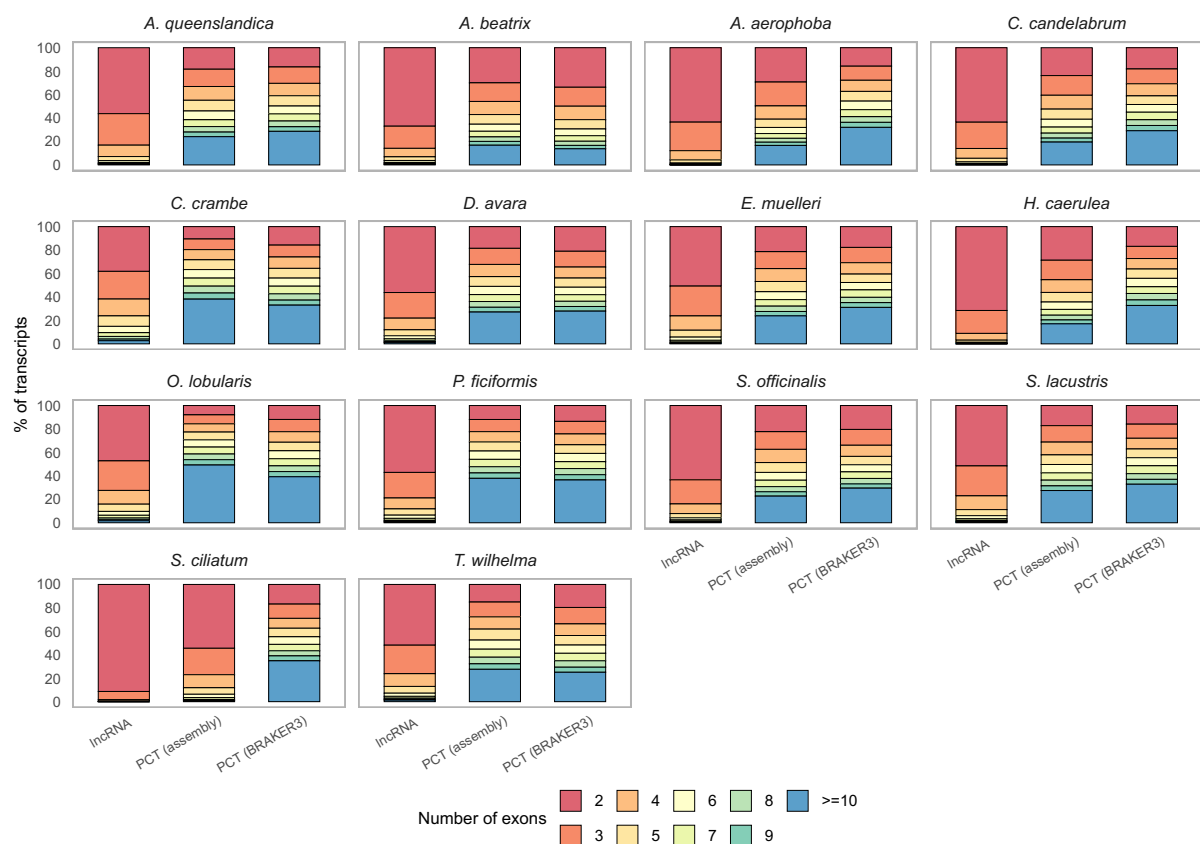


Figure 54. Distribution of exon numbers in poly-exonic transcripts among identified lncRNAs, assembly-derived PCTs, and BRAKER3-derived PCTs across the analysed sponge species. lncRNA = long non-coding RNA, PCT = protein-coding transcript.

The distribution of isoform number per lncRNA genes and PCGs was broadly similar across the analysed sponge species (Figure 55). In most taxa, lncRNA loci exhibited significantly fewer isoforms per gene than assembly-derived PCGs ($p < 0.001$). The majority of lncRNA genes produced only one or two isoforms, with isoform counts rarely exceeding 10 transcripts per locus. In contrast, assembly-derived PCGs displayed substantially greater isoform diversity, frequently producing multiple transcript variants and reaching maxima of 157 isoforms in *A. aerophoba* and 228 isoforms in *H. caerulea*. An exception was observed in *S. ciliatum*, where

assembly-derived PCGs exhibited isoform distributions similar to those of lncRNA genes. BRAKER3-derived PCGs consistently exhibited lower isoform diversity than assembly-derived PCGs in all analysed species ($p < 0.001$), with the exception of *S. ciliatum*. In most taxa, BRAKER3 gene models were strongly enriched for single-isoform genes and rarely exceeded five to seven transcript variants per locus.

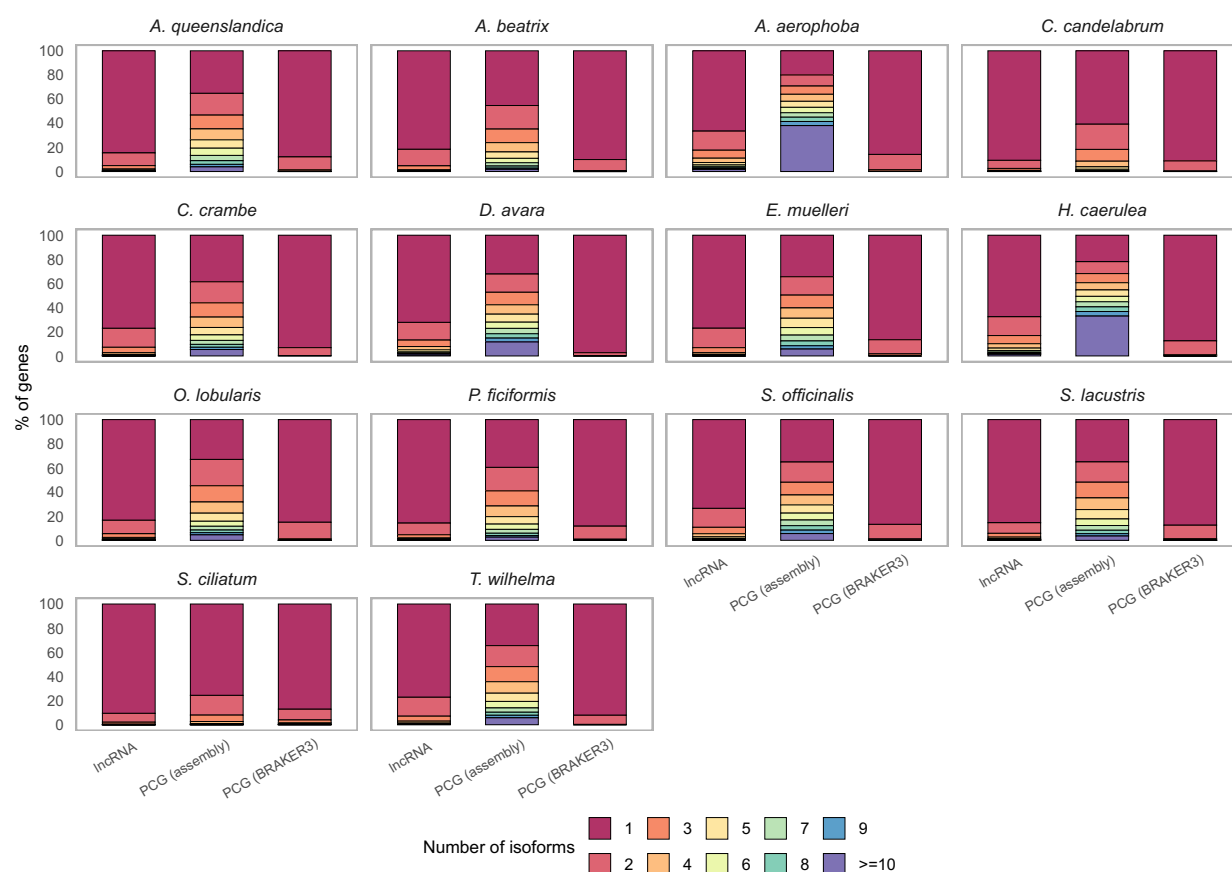


Figure 55. Distribution of isoform numbers per gene among identified lncRNA genes, assembly-derived PCGs, and BRAKER3-derived PCGs across the analysed sponge species. lncRNA = long non-coding RNA, PCG = protein-coding gene.

4.5.3. Long Non-coding RNA Splicing Analysis

Comparative splicing analysis was performed on introns derived from identified lncRNAs, assembly-derived PCGs, and BRAKER3-derived PCGs. Only splice junctions supported by read evidence were included in the analysis. Overall, the proportion of supported splice junctions was substantially lower in lncRNA introns (46.9%) compared to PCG introns from transcriptome assemblies (87.2%) and BRAKER3 PCG introns (90.8%).

Among confirmed splice junctions, the canonical GT-AG dinucleotide pair predominated across all transcript classes and sponge species, typically accounting for more than 95% of splice sites in assembly-derived and BRAKER3-derived PCG introns (Figure 56). In contrast, lncRNA introns exhibited a modest reduction in canonical GT-AG splice sites in several species,

accompanied by a relative enrichment of the non-canonical GC–AG and AT–AC splice junctions. GC–AG splice sites were generally rare in assembly-derived and BRAKER3-derived PCG introns, with median proportions of 1.0% and 0.9%, respectively, and reached a maximum of 4.8% in *O. lobularis*. Notably, GC–AG splice sites were significantly enriched in lncRNA introns from *S. ciliatum* (6.3%, $p < 0.001$) and *H. caerulea* (2.1%, $p = 0.019$) relative to both assembly-derived and BRAKER3-derived PCG introns. Similarly, AT–AC splice junctions accounted for less than 0.1% of splice sites in both protein-coding datasets, whereas lncRNA introns displayed substantially elevated proportions in several species. The highest proportions of AT–AC splice junctions were observed in lncRNA introns from *S. ciliatum* (7.6%), *H. caerulea* (4.5%), and *A. beatrix* (2.7%). In all three species, the frequency of AT–AC introns was significantly higher in lncRNA genes than in both assembly- and BRAKER3-derived PCG introns ($p < 0.001$).

MaxEntScan scores revealed clear differences in splice-site strength among splice junction classes (Figure 57). Analysed values represent the mean of the donor and acceptor splice-site scores calculated for each splice junction. Canonical GT–AG splice sites consistently exhibited the highest scores across all species and transcript classes, with median values typically ranging from 5.5 to 8.2. GT–AG splice-site strengths were highly similar among lncRNA, assembly-derived PCG, and BRAKER3-derived PCG introns in all analysed species. On the other hand, GC–AG splice sites displayed intermediate MaxEntScan scores, consistently lower than canonical GT–AG junctions but substantially higher than AT–AC splice sites. In assembly- and BRAKER3-derived PCG introns, GC–AG splice sites generally exhibited median scores between 4.0 and 5.4. In contrast, lncRNA introns frequently exhibited reduced GC–AG splice-site strength and greater score variability relative to both PCG classes, particularly in *H. caerulea* (median = 2.01), *D. avara* (median = 3.0), *S. officinalis* (median = 3.0), and *T. wilhelma* (median = 3.1), where all pairwise comparisons were significant ($p < 0.001$). In *S. ciliatum*, lncRNA introns displayed substantially higher MaxEntScan scores (median = 5.42) than both assembly-derived PCG (median = 3.66, $p < 0.001$) and BRAKER3-derived PCG introns (median = 1.77, $p < 0.001$). Nevertheless, comparatively strong GC–AG splice signals were retained in some lncRNA datasets, most notably in *E. muelleri* (median = 5.25).

Furthermore, AT–AC splice sites consistently produced strongly negative MaxEntScan scores across all datasets. Median AT–AC scores in lncRNA introns ranged from –6.4 to –12.4, with the weakest splice signals observed in *S. ciliatum* (–12.38) and *A. beatrix* (–12.23). When compared to lncRNA introns, assembly- and BRAKER3-derived PCG introns generally exhibited slightly higher AT–AC scores ($p < 0.001$), although these remained substantially lower than those of GT–AG and GC–AG splice sites. Notably, several species exhibiting pronounced enrichment of AT–AC splice junctions in lncRNA introns, including *A. beatrix*, *H. caerulea*, and *S. ciliatum*, also displayed some of the weakest AT–AC splice-site strengths.

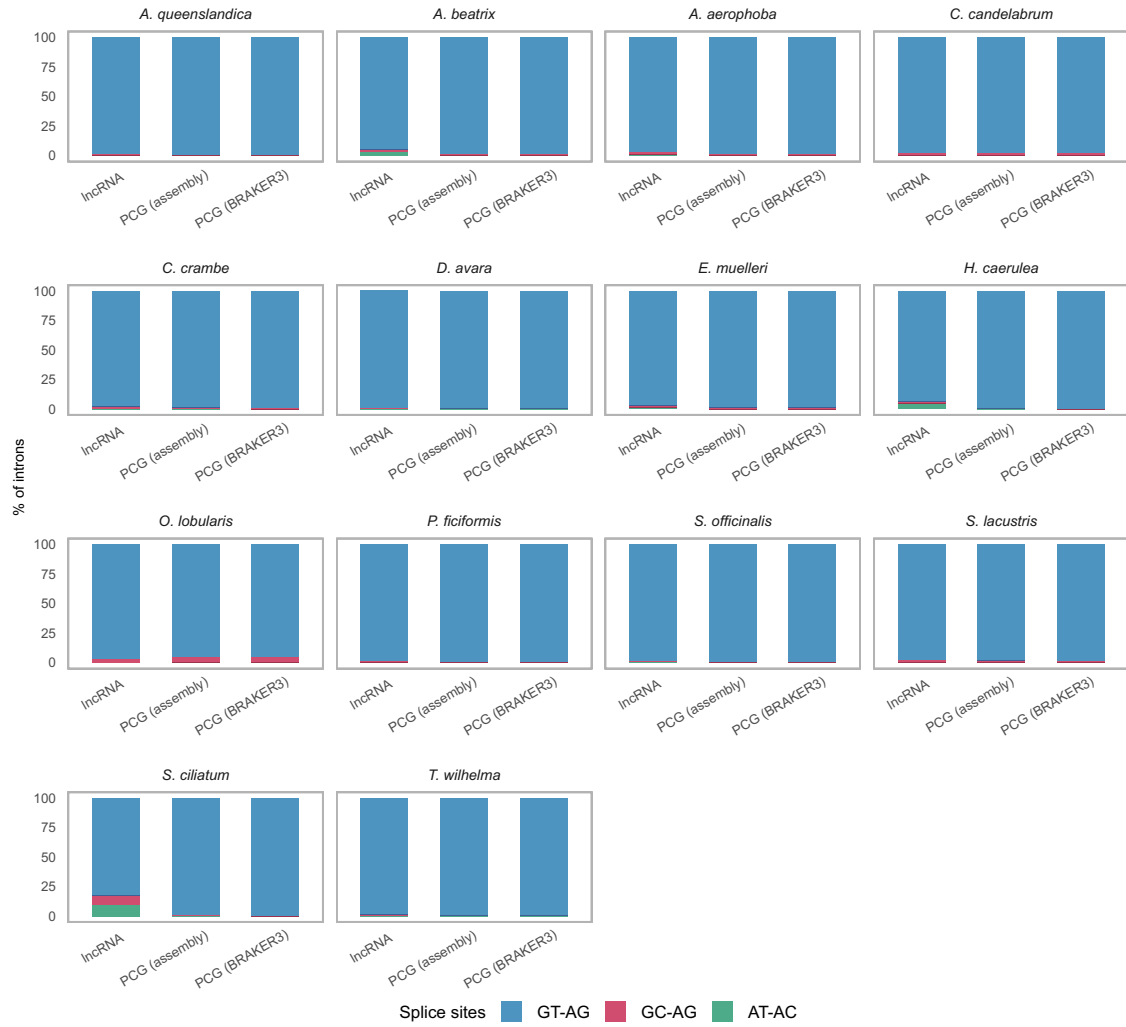


Figure 56. Splice-site composition of identified lncRNA introns, assembly-derived PCG introns, and BRAKER3-derived PCG introns across the analysed sponge species. lncRNA = long non-coding RNA, PCG = protein-coding gene.

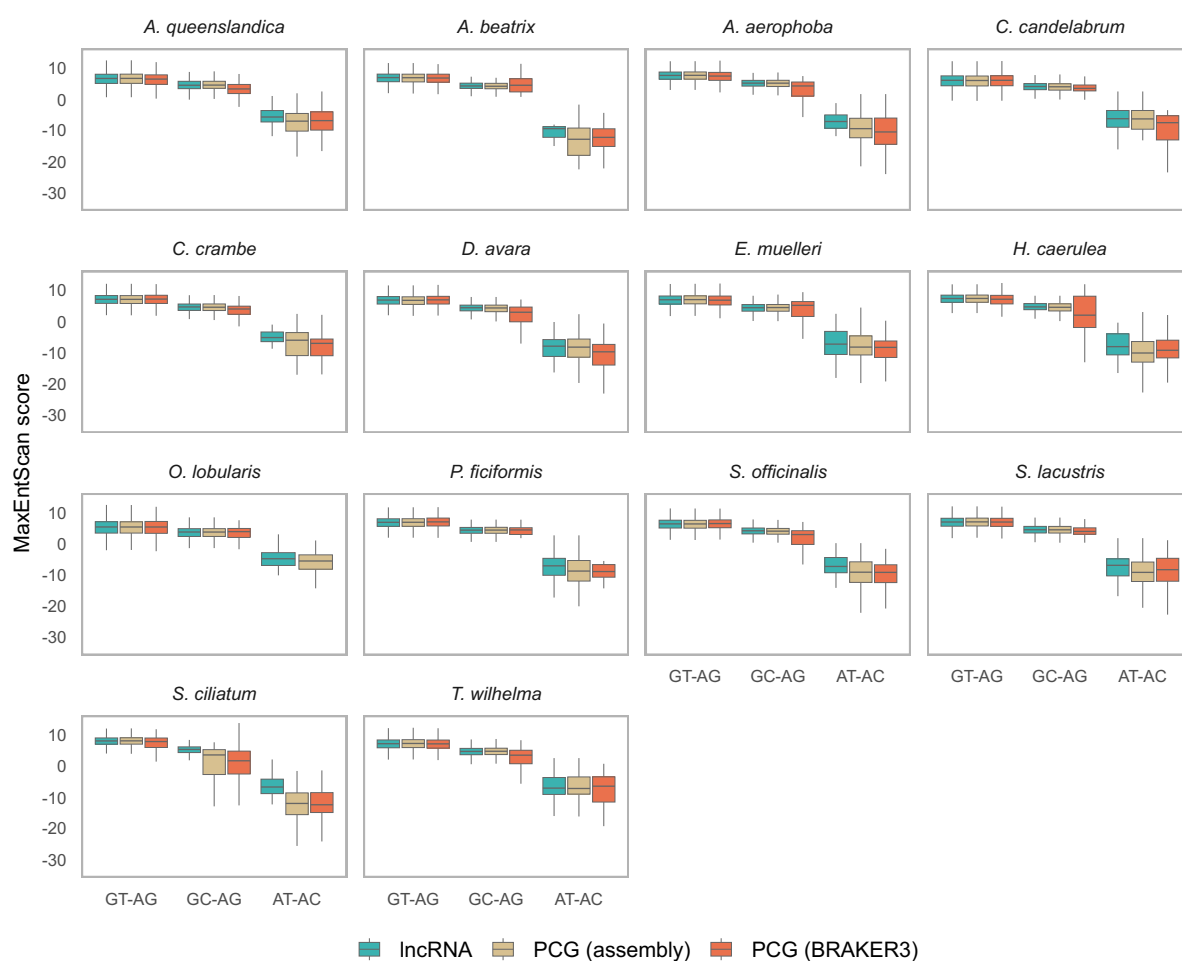


Figure 57. Distribution of mean MaxEntScan splice-site donor and acceptor scores across splice junctions in lncRNA introns, assembly-derived PCG introns, and BRAKER3-derived PCG introns across the analysed sponge species. lncRNA = long non-coding RNA, PCG = protein-coding gene.

Sequence logo analysis of canonical GT–AG splice junctions revealed highly similar donor and acceptor motifs across all transcript classes (Figure 58). Logos were generated using splice junctions aggregated across all analysed sponge species. Donor motifs exhibited strong conservation of the upstream AG dinucleotide together with a pronounced adenine-rich region immediately downstream of the donor site. Likewise, acceptor motifs displayed a pyrimidine-rich region upstream of the splice junction.

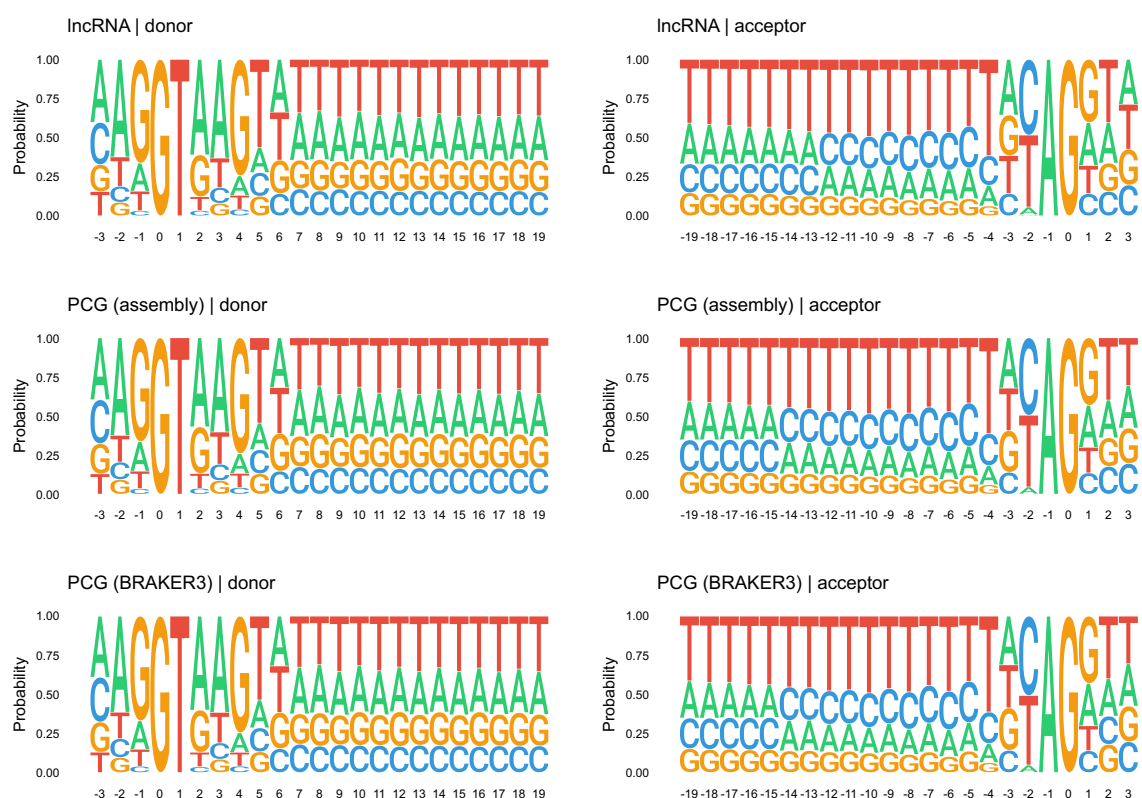


Figure 58. Sequence logos of canonical GT–AG splice junctions across lncRNA introns, assembly-derived PCG introns, and BRAKER3-derived PCG introns. Donor and acceptor splice-site motifs were generated using splice junctions aggregated across all sponge species. lncRNA = long non-coding RNA, PCG = protein-coding gene.

4.5.4. Long Non-coding RNA Sequence Composition

Next, sequence composition was compared among lncRNAs, assembly-derived PCTs, and BRAKER3-derived PCTs. Across the majority of analysed sponge species, UMAP projections of principal components derived from transcript k-mer frequencies revealed clear separation among the three transcript classes (Figure 59). In most species, lncRNAs formed distinct clusters that were clearly separated from both assembly-derived and BRAKER3-derived PCTs. The two PCT classes were generally more similar to one another, although partial separation was still evident in several species. The strongest separation between lncRNAs and PCTs was observed in *A. beatrix*, *A. queenslandica*, *C. crambe*, *H. caerulea*, *O. lobularis*, *S. officinalis*, and *T. wilhelma*. Notably, in *S. ciliatum*, assembly-derived PCTs and lncRNAs exhibited near-complete overlap, whereas BRAKER3-derived PCTs formed a clearly distinct cluster.

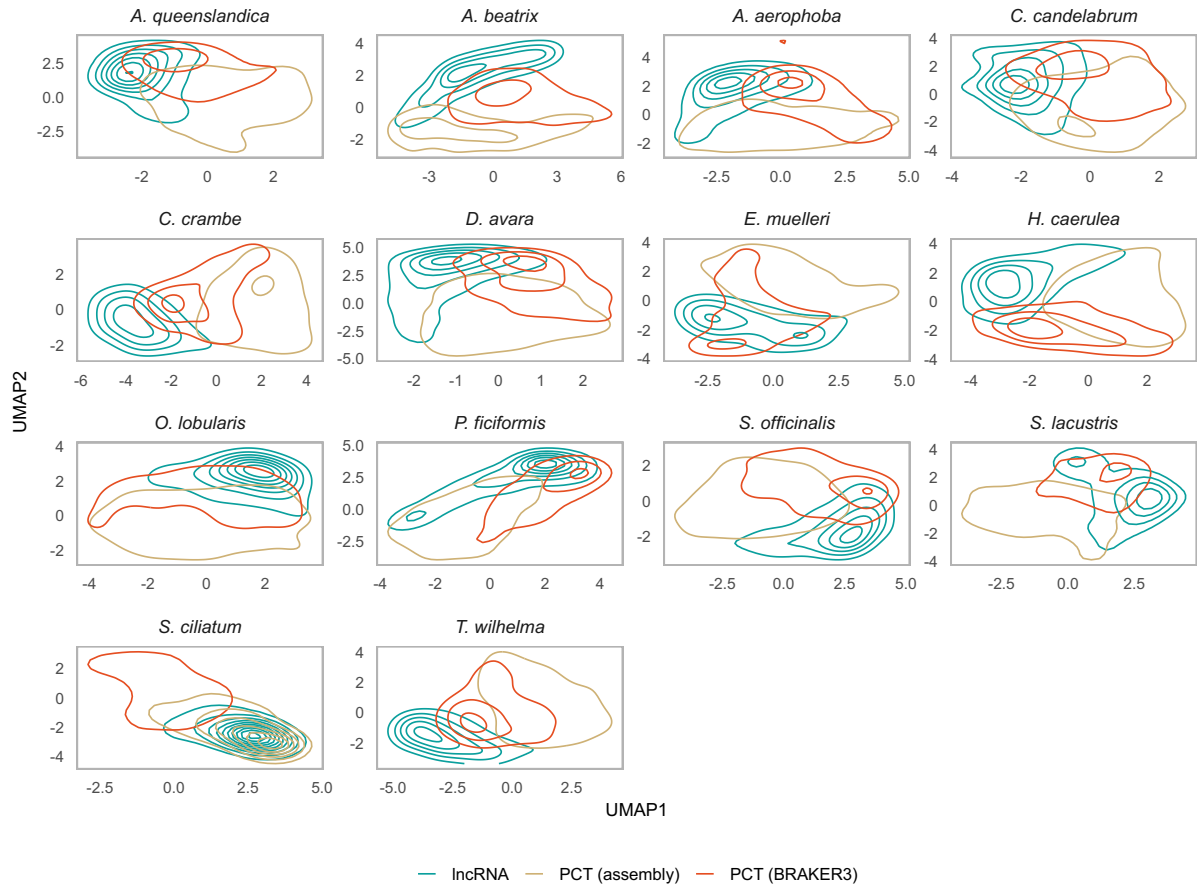


Figure 59. UMAP projection of principal components derived from transcript k-mer frequencies for lncRNAs, assembly-derived PCTs and BRAKER3-derived PCTs across the analysed sponge species. Density contours represent the distribution of transcript classes within reduced dimensional sequence composition space. lncRNA = long non-coding RNA, PCT = protein-coding transcript.

GC content distributions differed among lncRNAs, assembly-derived PCTs, and BRAKER3-derived PCTs across all analysed sponge species (Figure 60; all $p < 0.001$), although the magnitude of these differences was generally small. In most species, lncRNAs exhibited slightly lower GC content than PCTs, assembly-derived PCTs showed intermediate values, and BRAKER3-derived PCTs displayed marginally higher GC content. In contrast, GC content distributions were highly similar among transcript classes in *C. crambe*, *E. muelleri*, *O. lobularis*, and *T. wilhelma*.

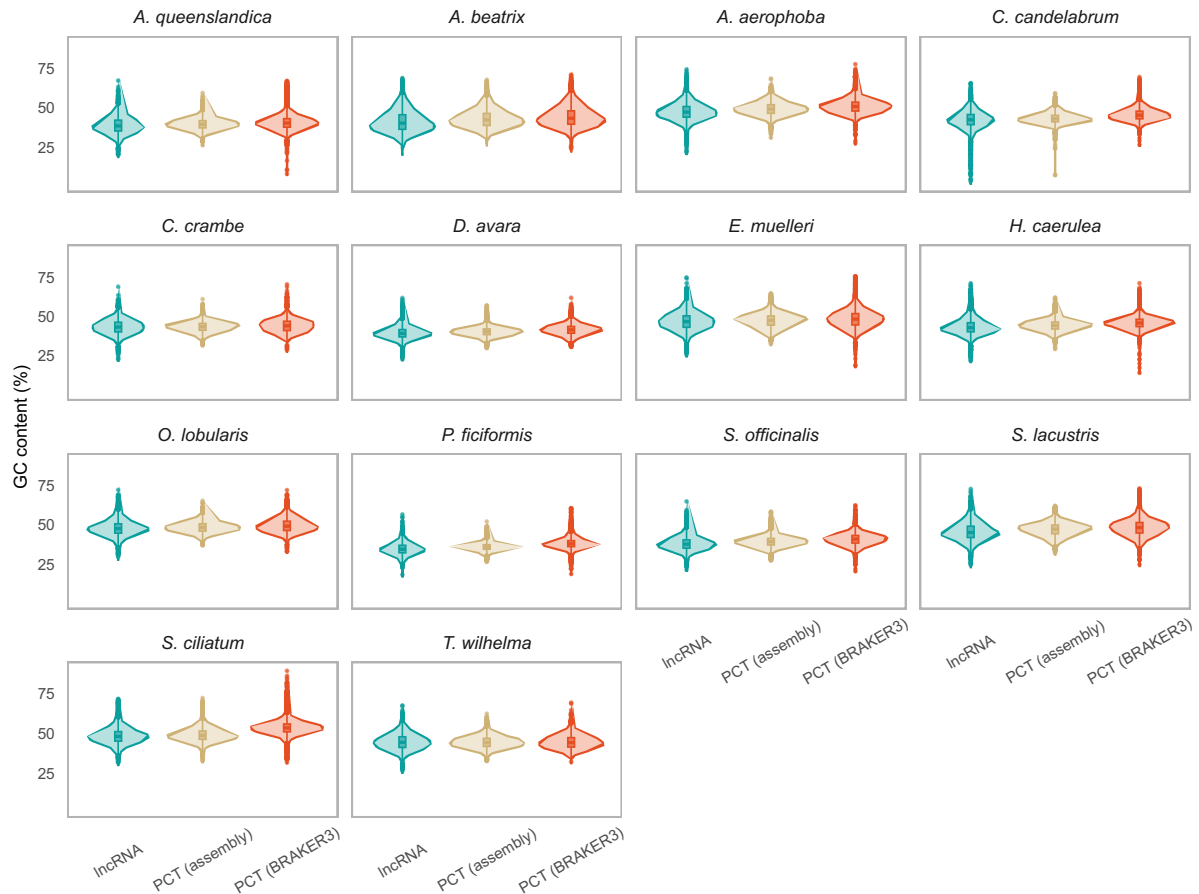


Figure 60. GC content distributions of lncRNAs, assembly-derived PCTs and BRAKER3-derived PCTs across the analysed sponge species. Violin plots show the distribution of transcript GC content, with embedded boxplots indicating median and interquartile range. lncRNA = long non-coding RNA, PCT = protein-coding transcript.

4.5.5. Genomic Organization of Long Non-coding RNAs Relative to Protein-coding Genes

The distribution of lncRNA genomic classes varied substantially among the analysed sponge species, although several consistent patterns were apparent across taxa (Figure 61). Intergenic lncRNAs represented the dominant class in most species, accounting for more than 50% of identified lncRNAs in *A. aerophoba*, *C. candelabrum*, *D. avara*, *O. lobularis*, *P. ficiformis*, *S. officinalis*, *S. lacustris*, and *T. wilhelma*. In contrast, antisense lncRNAs constituted the largest fraction in *A. beatrix* and *E. muelleri*. Intronic lncRNAs generally represented a smaller proportion of transcripts, although they were notably enriched in *H. caerulea* and *S. ciliatum*, where they accounted for a substantial fraction of identified lncRNAs. Overlapping lncRNAs were detected across all species but typically comprised a relatively minor component of the lncRNA repertoire, with the highest proportions observed in *A. aerophoba*, *H. caerulea*, and *S. ciliatum*. Bidirectional lncRNAs consistently represented the rarest category in all analysed species, usually accounting for less than 5% of transcripts.

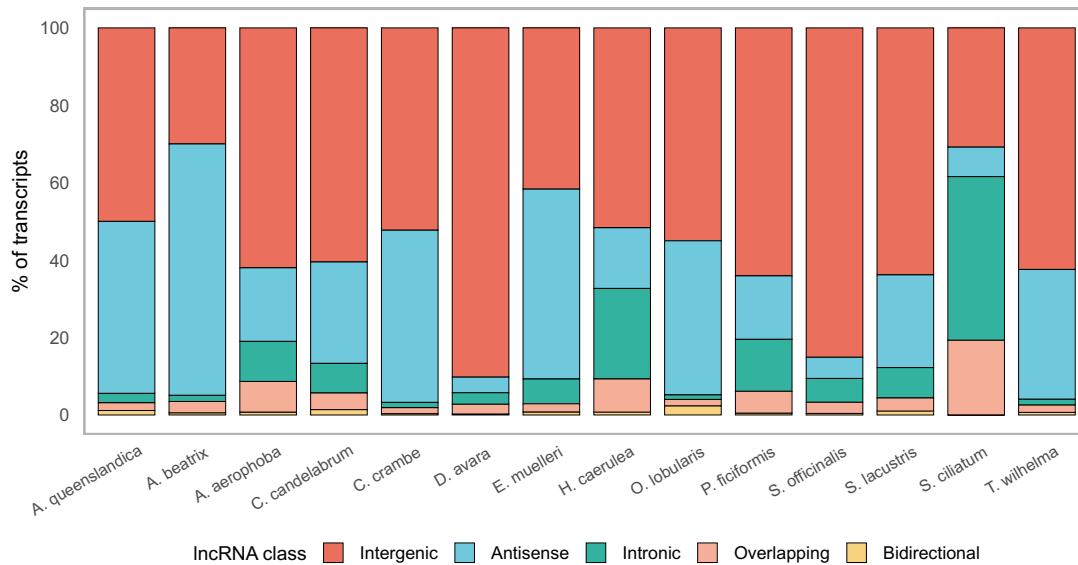


Figure 61. Genomic classification of lncRNAs relative to neighbouring PCGs across the analysed sponge species. lncRNA = long non-coding RNA.

Length distributions of lncRNA genomic elements showed moderate variation among genomic classes across the analysed sponge species (Figure 62). Antisense and overlapping lncRNAs generally possessed longer transcripts, with particularly long antisense transcripts observed in *A. beatrix* (median 693 bp) and *H. caerulea* (median 619 bp), and long overlapping transcripts identified in *A. beatrix* (median 651 bp), *P. ficiformis* (median 523 bp), and *T. wilhelma* (median 525 bp). In contrast, intronic and bidirectional lncRNAs were consistently shorter, with transcript lengths predominantly ranging between 330–445 bp across species. Intergenic lncRNAs occupied an intermediate position, with median transcript lengths ranging from 395–569 bp. Although the magnitude of these differences was generally modest, variation among antisense, intergenic, intronic, and overlapping lncRNA classes was statistically significant across all analysed species ($p < 0.001$).

Gene model lengths exhibited the most pronounced and consistent differences among lncRNA biotypes. Overlapping lncRNAs displayed substantially longer gene bodies than all other classes across the vast majority of species ($p < 0.001$), with median lengths ranging from 3511 bp in *C. candelabrum* to 7084 bp in *C. crambe*, and exceeding 5000 bp in *A. queenslandica*, *A. beatrix*, *D. avara*, *O. lobularis*, and *S. officinalis*. A notable exception was observed in *S. ciliatum*, where overlapping lncRNAs exhibited similar gene lengths relative to other classes (median 1099 bp). Intergenic and antisense lncRNAs showed intermediate gene body lengths, with medians generally ranging from 574 to 1519 bp, while intronic and bidirectional lncRNAs were the most compact, with median gene lengths predominantly below 700 bp across species.

Exon lengths represented the most conserved feature across lncRNA classes, with median values ranging from 101 to 222 bp. No lncRNA class exhibited consistently distinct exon

lengths relative to the others, and interspecific variation was generally limited. A modest elevation in exon lengths was observed in *C. candelabrum*, where intronic lncRNAs displayed longer exons relative to other classes (median 154 bp, $p < 0.001$). Similarly, antisense and intronic lncRNAs displayed significantly longer exons in *P. ficiformis* (medians = 171 and 166 bp) and *S. officinalis* (medians = 189 and 192 bp) than other lncRNA classes within the same species ($p < 0.001$).

Intron lengths significantly varied across lncRNA biotypes. Overlapping lncRNAs consistently harboured longer introns than all other classes ($p < 0.001$), with the longest median intron lengths observed in *A. beatrix* (3458 bp), *A. aerophoba* (1655 bp), and *S. officinalis* (1595 bp). In contrast, intergenic lncRNAs from the same species exhibited considerably shorter introns, with median lengths ranging from 57 to 257 bp. Antisense, overlapping, and bidirectional lncRNAs also consistently exhibited short introns, with median lengths below 200 bp in the majority of species. An exception was observed in *S. ciliatum*, where intergenic and antisense lncRNAs displayed atypically long introns relative to the general pattern, with median lengths of 471 and 627 bp.

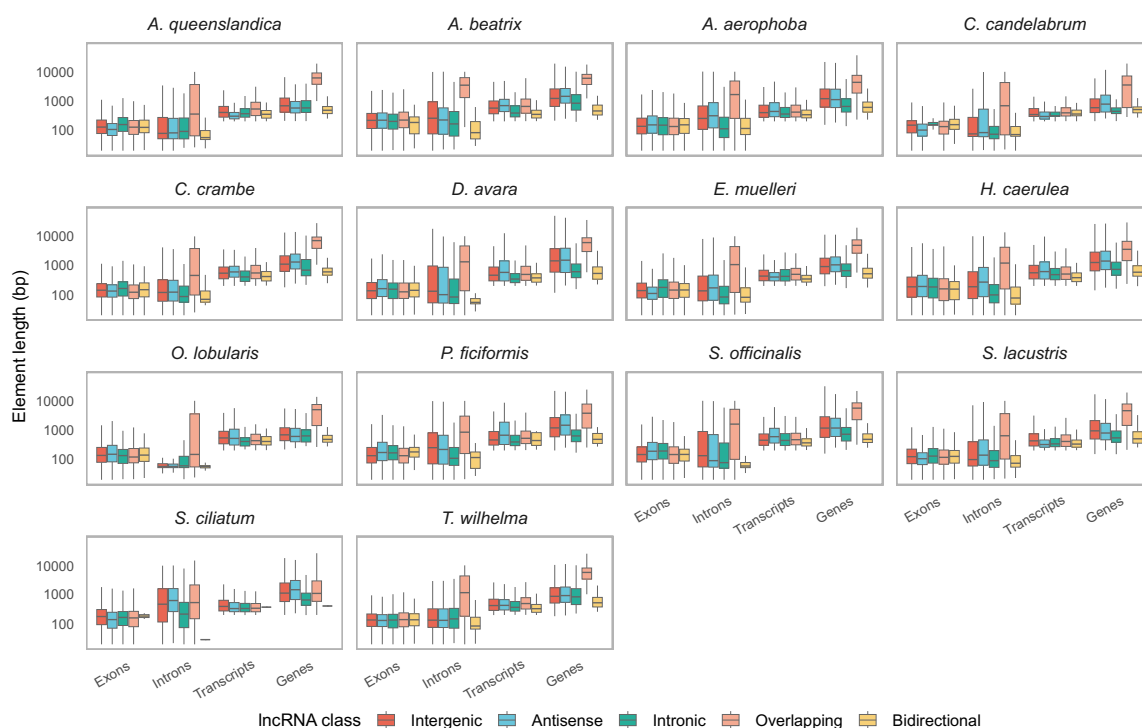


Figure 62. Distribution of exon, intron, transcript, and gene lengths among lncRNA classes across the analysed sponge species. lncRNA = long non-coding RNA.

Distribution of isoform numbers per gene varied moderately among lncRNA classes across sponge species (Figure 63). Overlapping lncRNAs exhibited the greatest transcript complexity in several species, with higher proportions of loci producing multiple isoforms relative to other classes. This trend was particularly pronounced in *A. queenslandica*, *C. crambe*, *E. muelleri*, *O. lobularis*, *S. officinalis*, and *T. wilhelma*. In contrast, intergenic and antisense lncRNAs generally

displayed similar isoform distributions, with only small fractions of loci generating more than three transcript isoforms. Notably, antisense lncRNAs exhibited reduced isoform diversity in *A. queenslandica*, *C. candelabrum*, *S. lacustris*, and *S. ciliatum* compared with other lncRNA classes ($p < 0.001$). Intronic lncRNAs generally showed isoform distributions comparable to those of intergenic and antisense lncRNAs, although reduced transcript complexity relative to other lncRNA classes was observed in *C. candelabrum*, *D. avara*, *O. lobularis*, *P. ficiformis*, *S. officinalis*, *S. lacustris*, and *S. ciliatum* ($p < 0.001$). Bidirectional lncRNAs generally displayed limited isoform diversity. However, *A. queenslandica*, *A. aerophoba*, *D. avara*, *E. muelleri*, *H. caerulea*, *P. ficiformis*, and *S. lacustris* exhibited a comparable proportion of bidirectional lncRNA loci producing multiple isoforms compared to other classes.

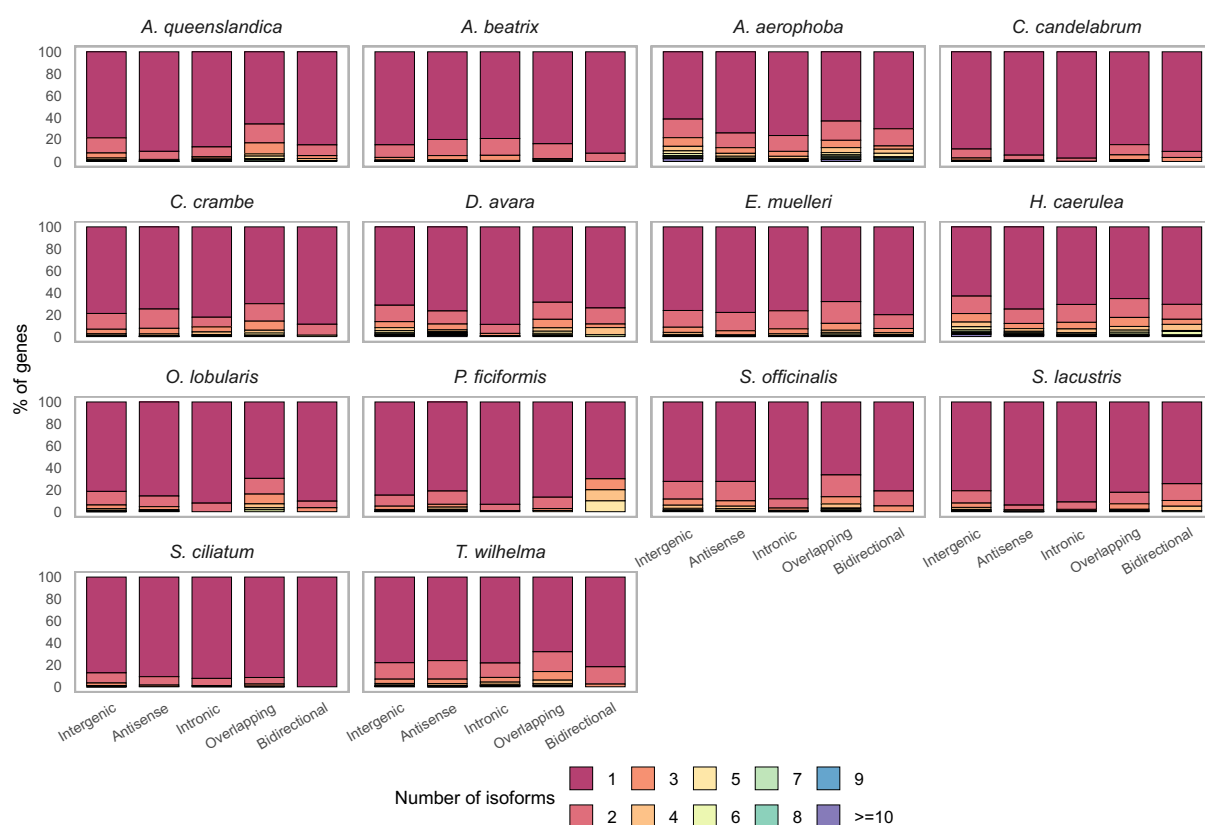


Figure 63. Distribution of isoform numbers among lncRNA classes across the analysed sponge species.

4.5.6. Association between Long Non-coding RNAs and Transposable Elements

Next, TE content was compared among lncRNA genes, assembly-derived PCGs, and BRAKER3-derived PCGs. First, each gene class was evaluated for the proportion of sequence overlapping selected TE classes. Across most analysed species, lncRNA genes contained a greater proportion of TE-derived sequence than both assembly-derived and BRAKER3-derived PCGs (Figure 64). For DNA transposons, lncRNA genes consistently showed higher relative TE content than PCGs in all species. Similarly, LTR retrotransposons displayed elevated relative sequence content within lncRNA genes in most species, particularly in *S. lacustris* and *S.*

ciliatum. Differences between gene classes were generally smaller for LINE elements, although lncRNA genes still tended to show moderately elevated relative TE content compared to PCGs in species such as *A. aerophoba*, *D. avara*, *S. officinalis*, and *S. ciliatum*. In contrast, SINEs and helitrons contributed only small fractions of total gene sequence across all species and categories, with minor differences between lncRNA genes and PCGs. Notable exceptions included increased SINE-associated sequence content of *S. ciliatum* and elevated helitron content in *C. candelabrum*.

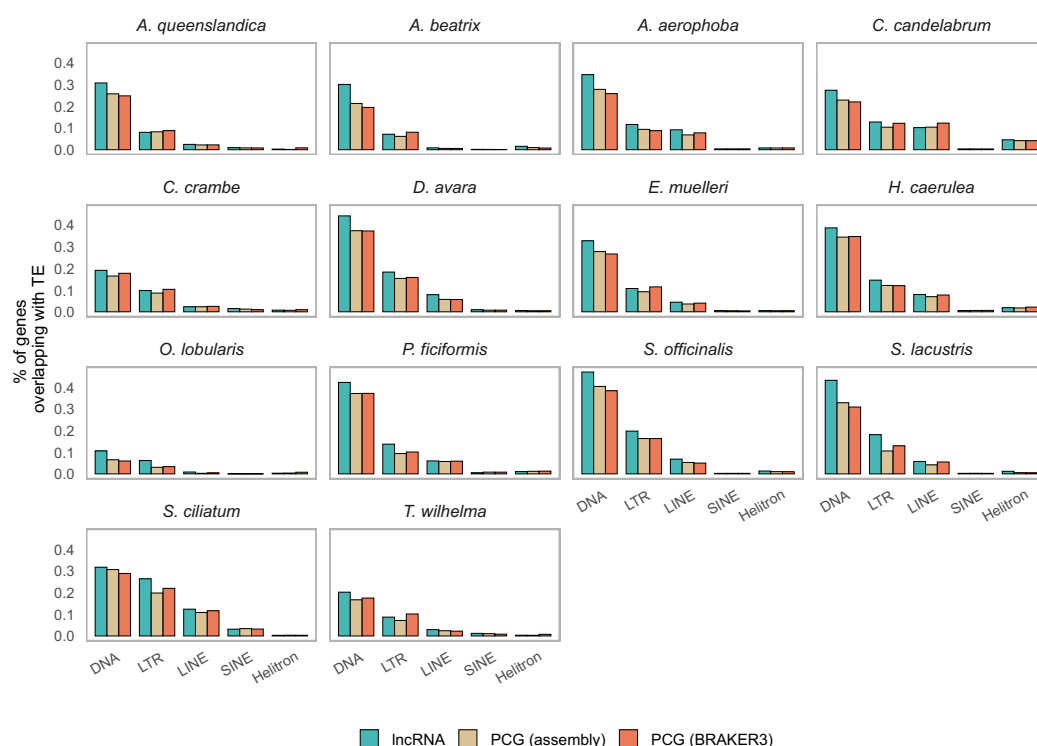


Figure 64. Relative TE content across lncRNA genes, assembly-derived PCGs and BRAKER3-derived PCGs. Bar plots show the proportion of gene sequence overlapping major TE classes. TE = transposable element, LTR = long terminal repeat retrotransposons, LINE = long interspersed nuclear elements, SINE = short interspersed nuclear elements, lncRNA = long non-coding RNA, PCG = protein-coding gene.

Exon-associated TE content showed patterns broadly similar to those observed for gene models, although overall proportions were generally lower (Figure 65). lncRNA exons consistently exhibited higher relative TE content than both assembly-derived and BRAKER3-derived protein-coding exons. Exonic lncRNA regions exhibited substantially higher relative DNA transposon content compared to protein-coding exons in the vast majority of species, with the exception of *C. crumbe*. Furthermore, lncRNA exons frequently showed elevated relative LTR content compared to protein-coding exons in seven species (*A. aerophoba*, *D. avara*, *H. caerulea*, *O. lobularis*, *P. ficiformis*, *S. officinalis*, and *S. ciliatum*). Differences between transcript classes were generally smaller for LINE elements, although lncRNA exons still tended to exhibit moderately increased TE content relative to protein-coding exons in

species such as *D. avara*, *H. caerulea*, *P. ficiformis*, *S. officinalis*, and *S. ciliatum*. Finally, SINEs and helitrons remained comparatively rare across exon regions in all transcript categories, with only modest differences among transcript classes. However, notable increases in helitron-associated exon content were observed in lncRNAs of *C. candelabrum* and *H. caerulea*.

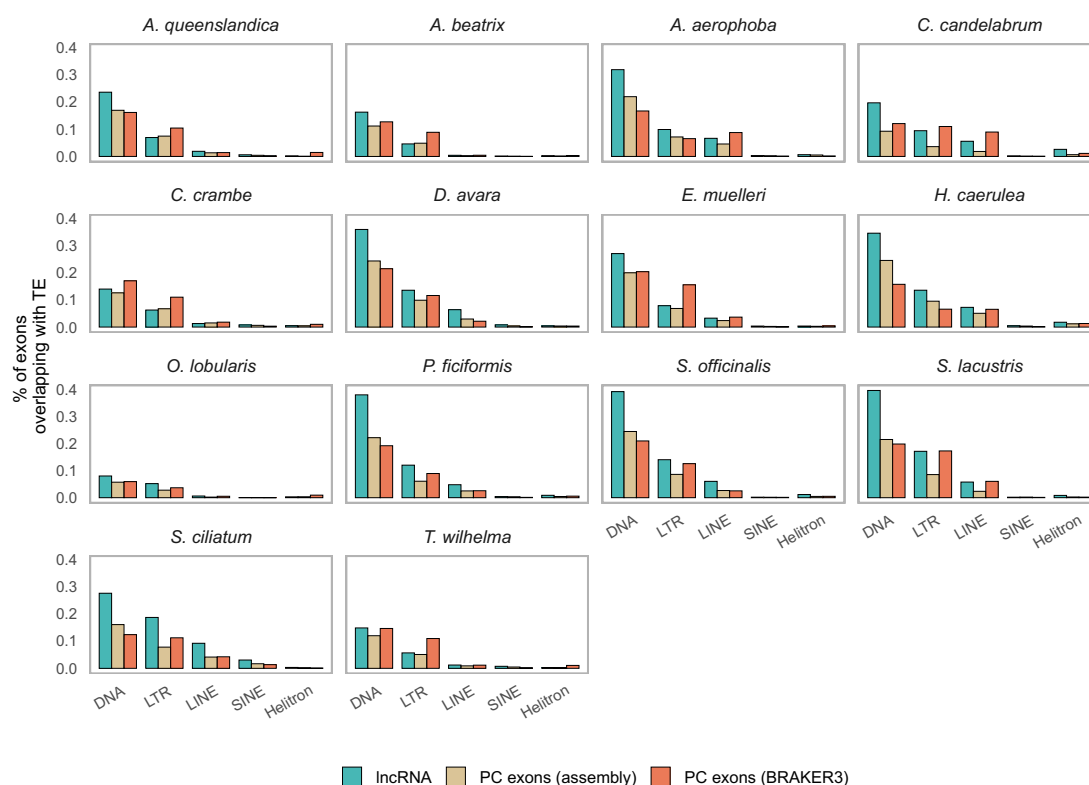


Figure 65. Relative TE content across lncRNA exons, assembly-derived protein-coding exons, and BRAKER3-derived protein-coding exons. Bar plots show the proportion of exonic sequence overlapping major TE classes. TE = transposable element, LTR = long terminal repeat retrotransposons, LINE = long interspersed nuclear elements, SINE = short interspersed nuclear elements, lncRNA = long non-coding RNA, PC = protein-coding.

Intron-associated TE content exhibited weaker and less consistent enrichment in lncRNA genes relative to PCGs than was observed for whole-gene or exonic regions (Figure 66). Although lncRNA introns generally displayed moderately higher DNA transposon content than PCG introns, the differences were typically smaller than those observed in exonic and whole-gene regions. LTR retrotransposon content was also broadly similar between lncRNA and PCG introns across most species, with only minor differences among transcript classes. However, markedly elevated intronic LTR content relative to PCG introns was observed in lncRNAs from *O. lobularis*, *P. ficiformis*, *S. lacustris*, and *S. ciliatum*. Comparable patterns were observed for LINEs, SINEs, and helitrons, which exhibited limited and inconsistent differences between lncRNA and PCG introns.

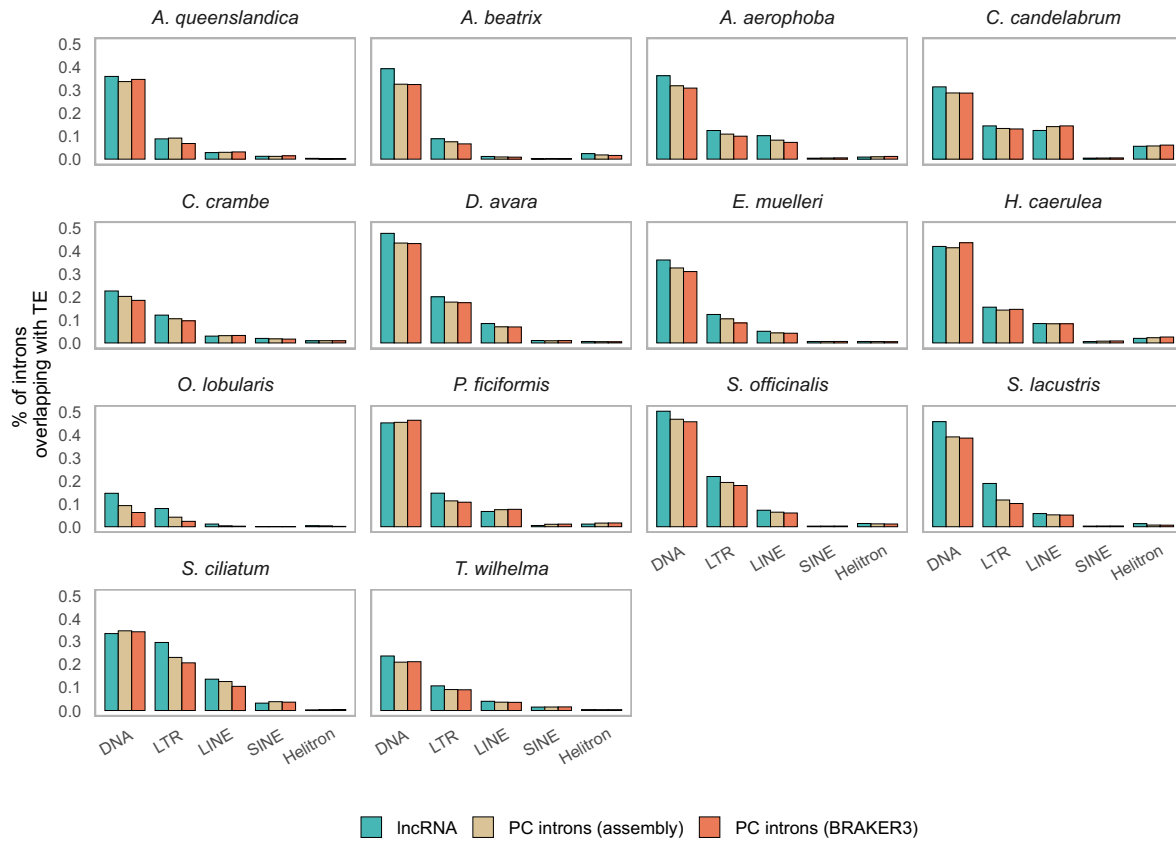


Figure 66. Relative TE content across IncRNA introns, assembly-derived PCG introns and BRAKER3-derived PCG introns. Bar plots show the proportion of intronic sequence overlapping major TE classes. TE = transposable element, LTR = long terminal repeat retrotransposons, LINE = long interspersed nuclear elements, SINE = short interspersed nuclear elements, IncRNA = long non-coding RNA, PC = protein-coding.

Next, TE enrichment was evaluated relative to the genome-wide abundance of each TE class in IncRNA genes, exons, and introns (Figure 67). Overall, IncRNA exons were generally depleted for most TE classes relative to genomic background, with enrichment values typically below 1. In contrast, IncRNA introns showed higher TE enrichment than exons across most species and TE classes. As expected, gene-level enrichment was intermediate between exon and intron values. DNA transposons generally showed enrichment values close to genomic background at the gene and intron levels, whereas exons were usually depleted. DNA transposon content exceeded the genomic background in introns of *A. queenslandica*, *A. aerophoba*, *C. candelabrum*, *O. lobularis*, and *S. ciliatum*. Similar patterns were observed for LTR retrotransposons, which were enriched in introns of *A. aerophoba*, *O. lobularis*, *P. ficiformis*, *S. lacustris*, and *S. ciliatum*. Furthermore, LINE elements were significantly enriched in introns of *C. candelabrum*, *C. crambe*, *O. lobularis*, and *T. wilhelma*. Interestingly, SINE elements exhibited elevated abundance within IncRNA introns in several species, with the strongest enrichment observed in *A. queenslandica* and *T. wilhelma*. Helitrons displayed the most variable pattern among TE classes. Intronic helitron enrichment was detected in *D. avara*

and *S. lacustris*, whereas depletion was observed in several other species, particularly *A. queenslandica*, *O. lobularis*, and *T. wilhelma*.

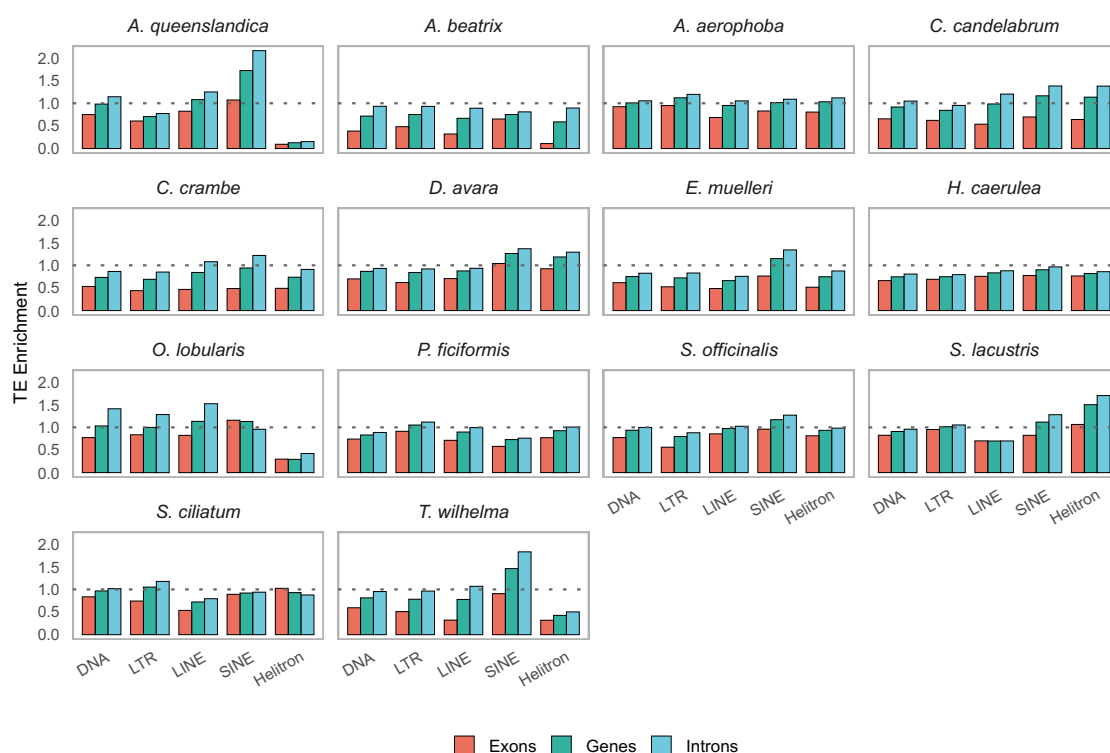


Figure 67. Enrichment of major TE classes in lncRNA genes, exons, and introns relative to their genome-wide abundance. Enrichment was calculated as the proportion of each genomic element class occupied by a given TE class divided by the genome-wide proportion of the TE class. TE = transposable element, LTR = long terminal repeat retrotransposons, LINE = long interspersed nuclear elements, SINE = short interspersed nuclear elements.

4.5.7. Long Non-coding RNA Evolutionary Conservation

Evolutionary conservation was evaluated across the detected lncRNAs in the analysed sponge species using sequence similarity and synteny-based approaches. Pairwise sequence conservation was quantified as the proportion of lncRNAs with high-confidence sequence matches between species. Because this metric was symmetrised by averaging reciprocal species comparisons, conservation values are reported as a single shared value for each species pair. Sequence-based comparisons revealed generally limited sharing of lncRNAs among the analysed sponge species (Figure 68). For most species pairs, fewer than 10% of lncRNAs exhibited detectable similarity. Nevertheless, several comparisons showed substantially higher levels of sequence sharing. The highest proportion of conserved lncRNAs was observed between the freshwater demosponges *E. muelleri* and *S. lacustris* (30.2%), followed by *C. crambe* and *T. wilhelma* (24.3%), and *D. avara* and *S. officinalis* (18.3%). Sequence similarity among conserved lncRNAs remained moderate. Median similarity scores, calculated as the product of sequence identity and alignment coverage, were 0.59 for *E.*

muelleri–*S. lacustris*, 0.60 for *C. crambe*–*T. wilhelma*, and 0.63 for *D. avara*–*S. officinalis* species pairs.

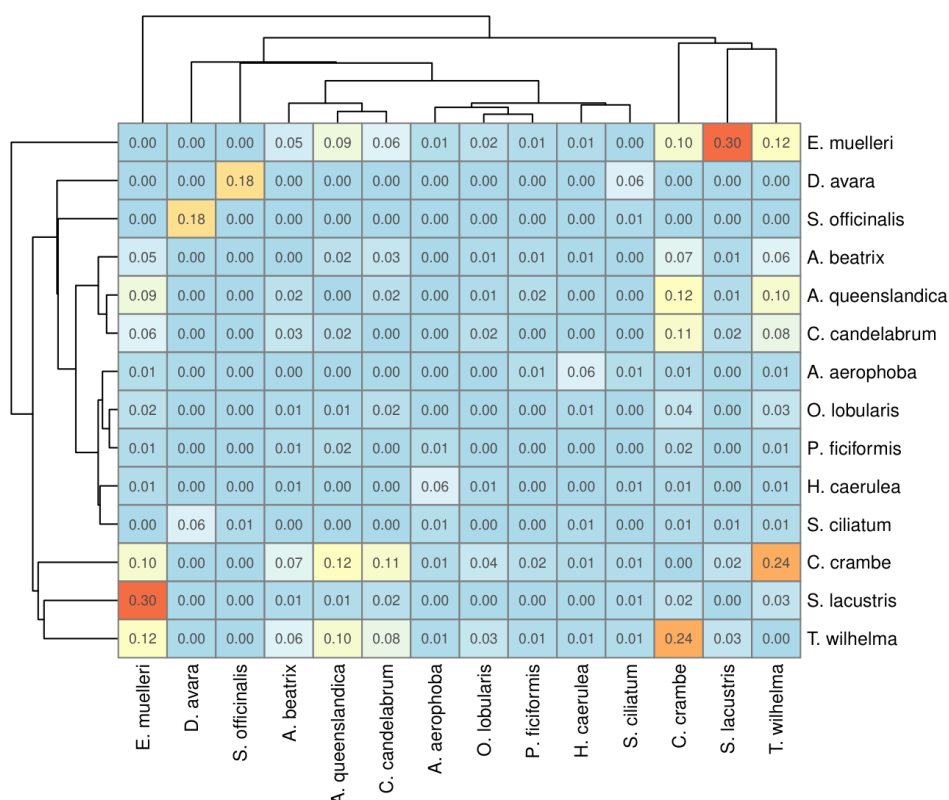


Figure 68. Proportion of shared lncRNAs across the analysed sponge species based on high-confidence sequence matches. Values represent pairwise sequence conservation estimates obtained by averaging reciprocal species comparisons.

Sequence conservation was further investigated by identifying interspecies lncRNA sequence similarity clusters. Analysis was restricted to clusters containing lncRNAs from at least three species. This yielded 1110 clusters spanning between three and eight species, with the majority comprising lncRNAs from only three or four species (Table 17). Mean alignment length increased with the number of species represented within a cluster, ranging from 1084 bp (SD = 742 bp, range = 297–9000 bp) in three-species clusters to 2139 bp (SD = 954 bp, range = 1464–2813 bp) in eight-species clusters. Similarly, mean alignment gap frequency increased from 0.31 (SD = 0.10) in three-species clusters to 0.52 (SD = 0.07) in eight-species clusters.

Table 17. Distribution of sequence-similarity lncRNA clusters across sponge species. Gap frequency represents the proportion of gap characters in the sequence alignment.

Species per cluster	Clusters N (%)	Alignment length mean (SD) (bp)	Length range (bp)	Gap frequency mean (SD)	Gap frequency range
3	673 (60.6)	1084 (742)	297–9000	0.31 (0.10)	0.06–0.63
4	294 (26.5)	1296 (760)	406–5786	0.42 (0.10)	0.20–0.68
5	102 (9.2)	1495 (947)	462–6465	0.45 (0.09)	0.31–0.76
6	30 (2.7)	1530 (646)	752–3639	0.48 (0.07)	0.34–0.67
7	9 (0.8)	2153 (1003)	1219–4395	0.50 (0.06)	0.36–0.73
8	2 (0.2)	2139 (954)	1464–2813	0.52 (0.02)	0.50–0.54

bp = base pair

Among the 1110 sequence-similarity clusters conserved across three or more species, 584 (52.6%) were composed entirely of lncRNAs assigned to the same genomic class. Class-conserved clusters were strongly enriched for antisense lncRNAs. Antisense transcripts accounted for 67.6% (395/584) of all class-conserved clusters despite representing only 26.3% of the total lncRNA repertoire, indicating a significant enrichment among lncRNA clusters ($p < 0.001$). In contrast, intergenic lncRNAs comprised 60.3% of all detected lncRNAs but only 36.3% (212/584) of class-conserved clusters, suggesting significant depletion relative to their genomic abundance ($p < 0.001$). Likewise, intronic, overlapping, and bidirectional lncRNAs were all significantly underrepresented among class-conserved clusters compared with their frequencies in the complete lncRNA dataset ($p < 0.001$).

To complement sequence-based analyses, lncRNA conservation was further investigated using a synteny-based approach. Conserved genomic blocks were identified from collinear PCG regions shared between species. Under the employed definition, lncRNA loci were considered syntenically conserved when they occurred within corresponding syntenic neighbourhoods and satisfied class-specific genomic context criteria. Specifically, antisense, intronic, and overlapping lncRNAs were required to be associated with orthologous PCGs, whereas intergenic and bidirectional lncRNAs were required to share conserved flanking gene context.

The proportion of lncRNAs with syntenic support in at least one other species ranged from 0.0% to 12.2% across species (Table 18). The highest proportions were observed in the freshwater demosponges *S. lacustris* (12.2%) and *E. muelleri* (9.0%). Intermediate levels were detected in *T. wilhelma* (7.5%), *C. crambe* (7.0%), and *D. avara* (4.9%), while most remaining species exhibited values below 5%. The lowest proportions were found in *C. candelabrum* (0.5%) and *O. lobularis* (0.8%). No synteny-supported lncRNA loci were detected in *A. beatrix* or *S. ciliatum*. The identified syntenic blocks contained on average between 3.4 and 4.2 conserved PCGs per block. Mean block lengths varied substantially, ranging from 90.3 kb in *O. lobularis* to 4551.3 kb in *D. avara*. Most synteny-supported lncRNAs were associated with blocks conserved in only one partner species.

Table 18. Summary statistics of synteny-supported lncRNAs across sponge species.

Species	Syntenic lncRNAs N (%)	Blocks N	Proteins/block mean (SD)	Block length mean (SD) (kb)	Species N mean (SD)
<i>A. queenslandica</i>	326 (2.8%)	205	3.6 (0.8)	121.0 (83.6)	1.2 (0.4)
<i>A. beatrix</i>	0 (0.0%)	/	/	/	/
<i>A. aerophoba</i>	772 (2.7%)	226	3.6 (0.8)	454.4 (314.7)	1.0 (0.1)
<i>C. candelabrum</i>	0 (0.0%)	34	3.6 (0.8)	210.8 (163.5)	1.0 (0.0)
<i>C. crambe</i>	1866 (7.0%)	429	3.7 (0.8)	328.4 (263.0)	1.2 (0.5)
<i>D. avara</i>	1814 (4.9%)	188	3.6 (0.8)	4551.3 (3416.2)	1.0 (0.2)
<i>E. muelleri</i>	3386 (9.0%)	494	3.9 (1.0)	1072.6 (1012.3)	1.1 (0.4)
<i>H. caerulea</i>	978 (3.1%)	232	3.7 (0.8)	512.3 (358.2)	1.0 (0.1)
<i>O. lobularis</i>	35 (0.8%)	30	4.2 (0.9)	90.3 (73.4)	1.0 (0.0)
<i>P. ficiformis</i>	153 (4.1%)	122	3.5 (0.8)	296.3 (173.0)	1.1 (0.3)
<i>S. officinalis</i>	659 (2.8%)	173	3.4 (0.6)	3133.4 (2427.6)	1.0 (0.2)
<i>S. lacustris</i>	1451 (12.2%)	298	4.0 (1.0)	1375.7 (1077.8)	1.1 (0.4)
<i>S. ciliatum</i>	0 (0.0%)	/	/	/	/
<i>T. wilhelma</i>	1652 (7.5%)	337	3.7 (0.9)	315.2 (234.9)	1.2 (0.5)

kb = kilobase

Pairwise synteny conservation was quantified as the proportion of lncRNAs in each species with at least one syntenic counterpart in another species. Overall, syntenic conservation of lncRNA loci was limited, with most pairwise comparisons yielding values close to zero (Figure 69). The highest levels of synteny were observed between the freshwater demosponges *E. muelleri* and *S. lacustris*, with 11.8% and 7.3% of lncRNAs possessing syntenic counterparts in the species pair. A secondary signal of conservation was detected between *C. crambe* and *T. wilhelma*, where pairwise values reached 4.1% and 5.4%.

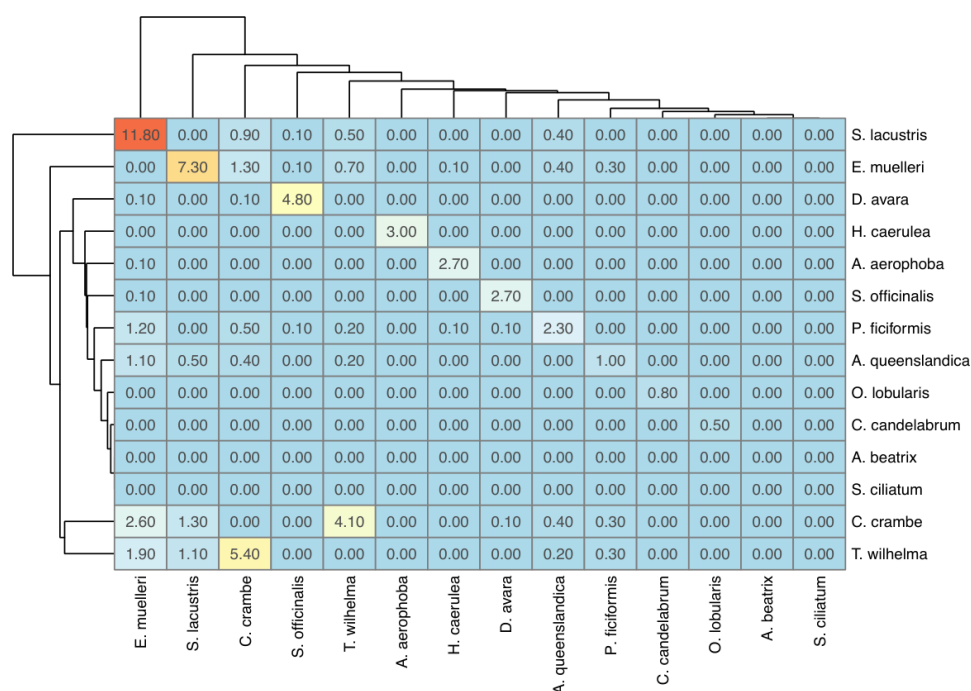


Figure 69. Pairwise synteny conservation of lncRNAs across the analysed sponge species. Cell values represent the proportion of lncRNAs in the query species that possess at least one syntenic counterpart in the subject species.

The genomic distribution of syntenic lncRNAs between *E. muelleri* and *S. lacustris* was visualised across the 15 longest *E. muelleri* chromosomes and their best syntenic counterparts in *S. lacustris* (Figure 70). Under the applied synteny criterion, 179 syntenic blocks were identified between the two species, encompassing 2754 *E. muelleri* and 1938 *S. lacustris* lncRNAs. Syntenic blocks were detected across all examined chromosomes, with syntenic lncRNAs distributed broadly throughout the genome. Several chromosome pairs exhibited crossing ribbon patterns. Among all syntenic lncRNA pairs, 302 exhibited direct sequence similarity between syntenic counterparts. These corresponded to 293 *E. muelleri* lncRNAs (10.6% of all *E. muelleri* lncRNAs with syntenic support in *S. lacustris*) and 300 *S. lacustris* lncRNAs (15.5%).

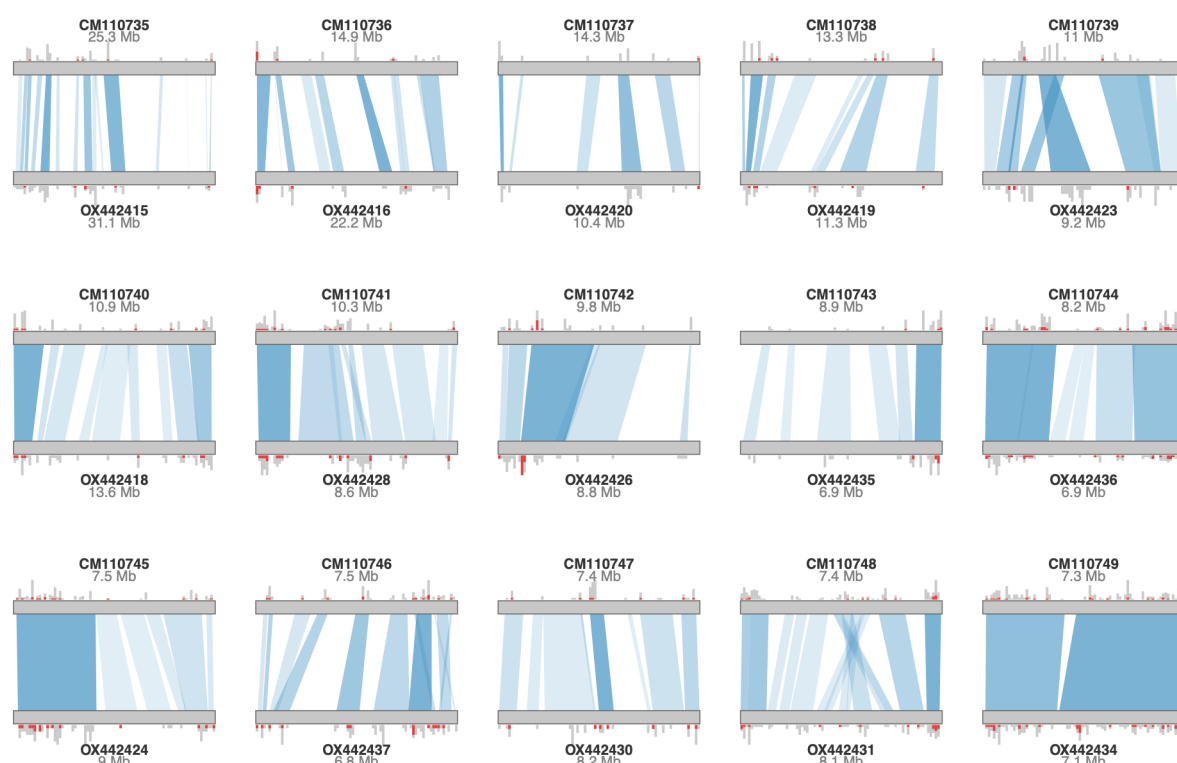


Figure 70. Chromosomal synteny and lncRNA conservation between *E. muelleri* and *S. lacustris*. Each panel depicts a pairwise comparison between one of the 15 longest *E. muelleri* chromosomes (top bar) and its best syntenic partner in *S. lacustris* (bottom bar), ordered by decreasing *E. muelleri* chromosome length. Blue ribbons represent syntenic blocks, with opacity proportional to the number of lncRNAs contained within each block. Density tracks show the genomic distribution of all syntenic lncRNAs (gray) and the subset with direct sequence similarity to their syntenic partner in the opposing species (red). Chromosome accession identifiers and lengths are indicated above and below each bar. Mb = megabase.

Under the employed synteny criterion, the maximum observed conservation breadth was five species, achieved by 12 lncRNA loci. These loci were distributed among *C. crambe* (nine loci), *S. lacustris* (two loci), and *T. wilhelma* (two loci). Expanding the analysis to lncRNAs conserved

across four or more species identified 169 broadly conserved loci spanning eight sponge species, with *C. crambe* (61 loci), *T. wilhelma* (45), and *E. muelleri* (41) contributing the largest numbers.

The most prominent conserved genomic neighbourhood was shared among five species (*C. crambe*, *D. avara*, *P. ficiformis*, *S. lacustris*, and *T. wilhelma*) and contained between seven (*D. avara* and *P. ficiformis*) and 11 (*C. crambe*) syntenic lncRNAs per species. This neighbourhood was defined by the conserved PCGs *ARMC2*, *FOXJ1*, *CFAP46*, *CBY1*, *LACE1*, and *FANCI*, most of which are associated with ciliogenesis. All lncRNAs within this conserved neighbourhood were intergenic, and none exhibited detectable sequence similarity to their syntenic counterparts.

Among the conserved syntenic regions, a TGFB3-associated neighbourhood was notable for its high density of conserved lncRNAs. In addition to TGFB3, the neighbourhood contained CHD, NVL, and RBM25, genes associated with chromatin regulation, ribosome biogenesis, and RNA splicing, respectively. This block was conserved across *A. aerophoba*, *H. caerulea*, *S. lacustris*, and *T. wilhelma*, with individual species contributing between four and 12 syntenic lncRNAs. Within this neighbourhood, lncRNAs were predominantly intergenic (51.1%), followed by antisense (44.2%) and intronic (3.5%) classes. Sequence similarity was detected in 6.6% of syntenic lncRNA pairs.

To investigate the biological functions associated with conserved lncRNAs, GO enrichment analysis was performed separately for PCGs linked to sequence-conserved, synteny-conserved, and sequence- and synteny-conserved lncRNAs. Within the sequence-conserved set, the most strongly enriched Biological Process terms were amide and peptide biosynthetic processes, both detected in *E. muelleri* and *S. lacustris* (Figure 71). Peptidyl-serine phosphorylation was enriched across four species (*C. crambe*, *E. muelleri*, *S. lacustris*, and *T. wilhelma*), while multi-organism reproductive process was enriched in *C. crambe*, *E. muelleri*, and *T. wilhelma*. Additional recurrent enrichments included epithelial cell development, detected in three species, and actin filament organisation, detected in two species. Several enriched terms originating from bilaterian model-organism annotations, including neuronal stem cell division, border follicle cell migration, and dorsal closure, were also identified.

The synteny-conserved set exhibited a distinct enrichment profile. Neuron axonogenesis and regulation of neuron apoptosis were each enriched in two species. Similar to the sequence-conserved set, peptidyl-serine phosphorylation was enriched in *E. muelleri* and *S. lacustris*. The sequence- and synteny-conserved set, comprising lncRNAs supported by both sequence and positional conservation, exhibited the most coherent enrichment profile. Peptidyl-serine phosphorylation and protein autophosphorylation were each enriched in three species, while ATF6-mediated unfolded protein response and protein folding in the endoplasmic reticulum were enriched in two species. Developmental specification terms, including muscle cell fate specification, endodermal cell fate specification, and midgut development, were likewise enriched in two species each. Notably, enrichment within the synteny-conserved and

sequence- and synteny-conserved sets was generally weaker than in the sequence-conserved set, with most enriched terms supported by relatively few genes (typically 2–4 per species) and exhibiting only modest statistical significance.

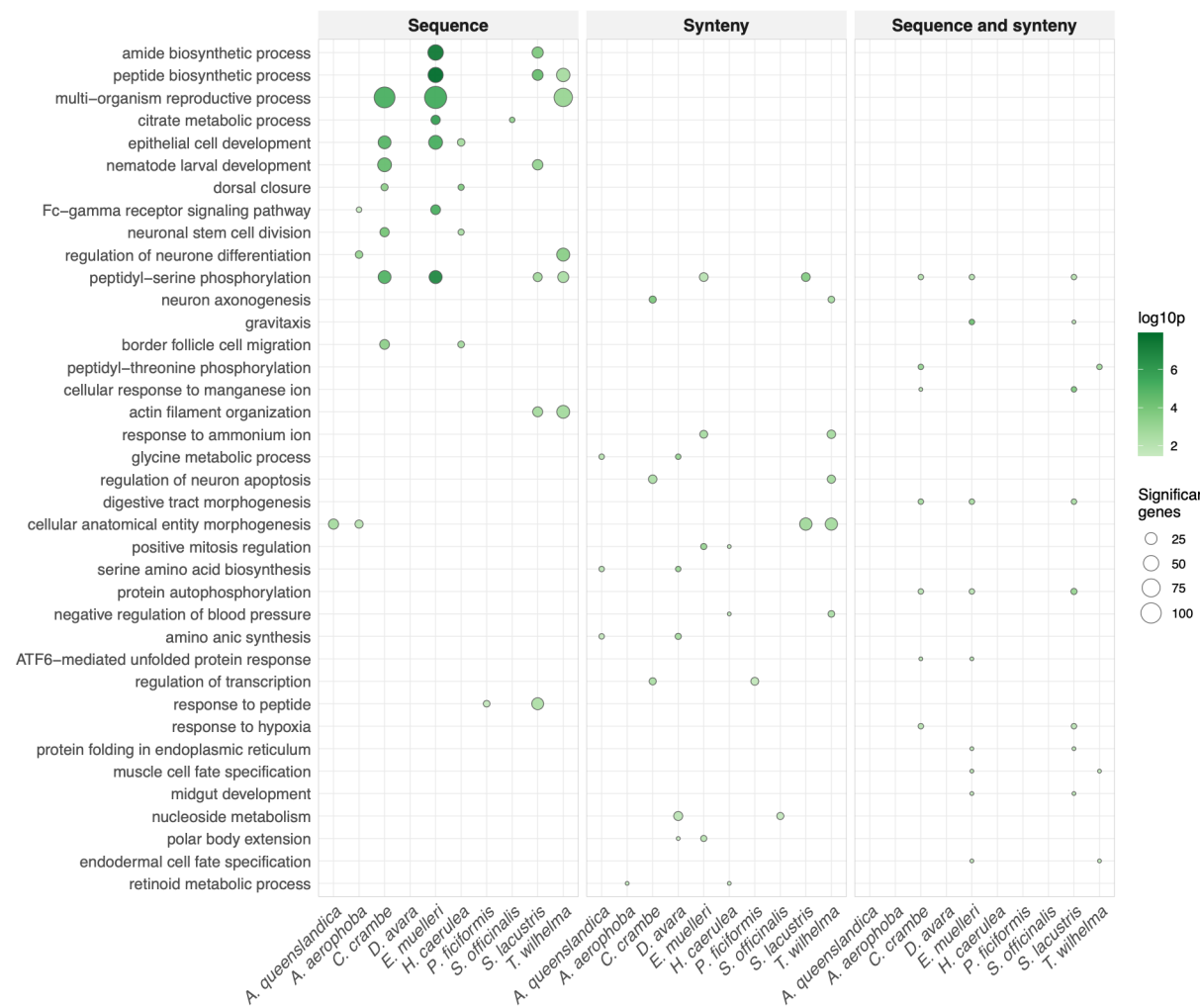


Figure 71. Biological Process GO enrichment associated with conserved lncRNA loci across the analysed sponge species.

Molecular Function GO enrichment of the PCGs associated with conserved lncRNA sets revealed a coherent pattern across all three conservation tiers (Figure 72). The most broadly enriched molecular function in the sequence-conserved set was ATP hydrolysis activity, detected in all seven species with significant enrichments (*A. queenslandica*, *A. aerophoba*, *C. crambe*, *E. muelleri*, *H. caerulea*, *S. lacustris*, and *T. wilhelma*). ATP binding was similarly enriched in five species. Dynein-associated functions were also prominent, with dynein light chain binding enriched in six species, and both dynein light intermediate chain binding and dynein intermediate chain binding enriched in three species. Additional enrichments for microfilament motor activity (three species), actin filament binding (two species), and microtubule motor activity (two species) further reinforced the recurring cytoskeletal and

ciliary theme. Protein folding chaperone activity was enriched in three species, calmodulin binding in two, and protein serine/threonine kinase activity in three species.

In the syntenic set, unfolded protein binding was the most recurrently enriched molecular function, detected in four species (*A. queenslandica*, *C. crambe*, *E. muelleri*, and *S. lacustris*). Furthermore, SH2 domain binding was enriched in three species, while calcium/calmodulin-dependent protein kinase activity and transcription coactivator activity were each enriched in two species. Phosphatase activity showed particularly strong enrichment in *H. caerulea*.

The double-conserved set exhibited the strongest enrichment for signal-transduction-related functions. For instance, SH2 domain binding was enriched in four species (*C. crambe*, *E. muelleri*, *S. lacustris*, and *T. wilhelma*), followed by MAP kinase activity in three species. Beta-catenin binding, calcium channel regulator activity, and histone kinase activity were each enriched in two species. As observed for Biological Process ontology, enrichment in the syntenic double-conserved sets was considerably weaker than in the sequence-conserved with most enriched Molecular Function terms involving only one significant gene per species and p-values generally above 0.01. Notably, no significantly enriched GO terms were detected in either *C. candelabrum* or *O. lobularis* due to the limited number of sequence- and synteny-conserved lncRNAs.

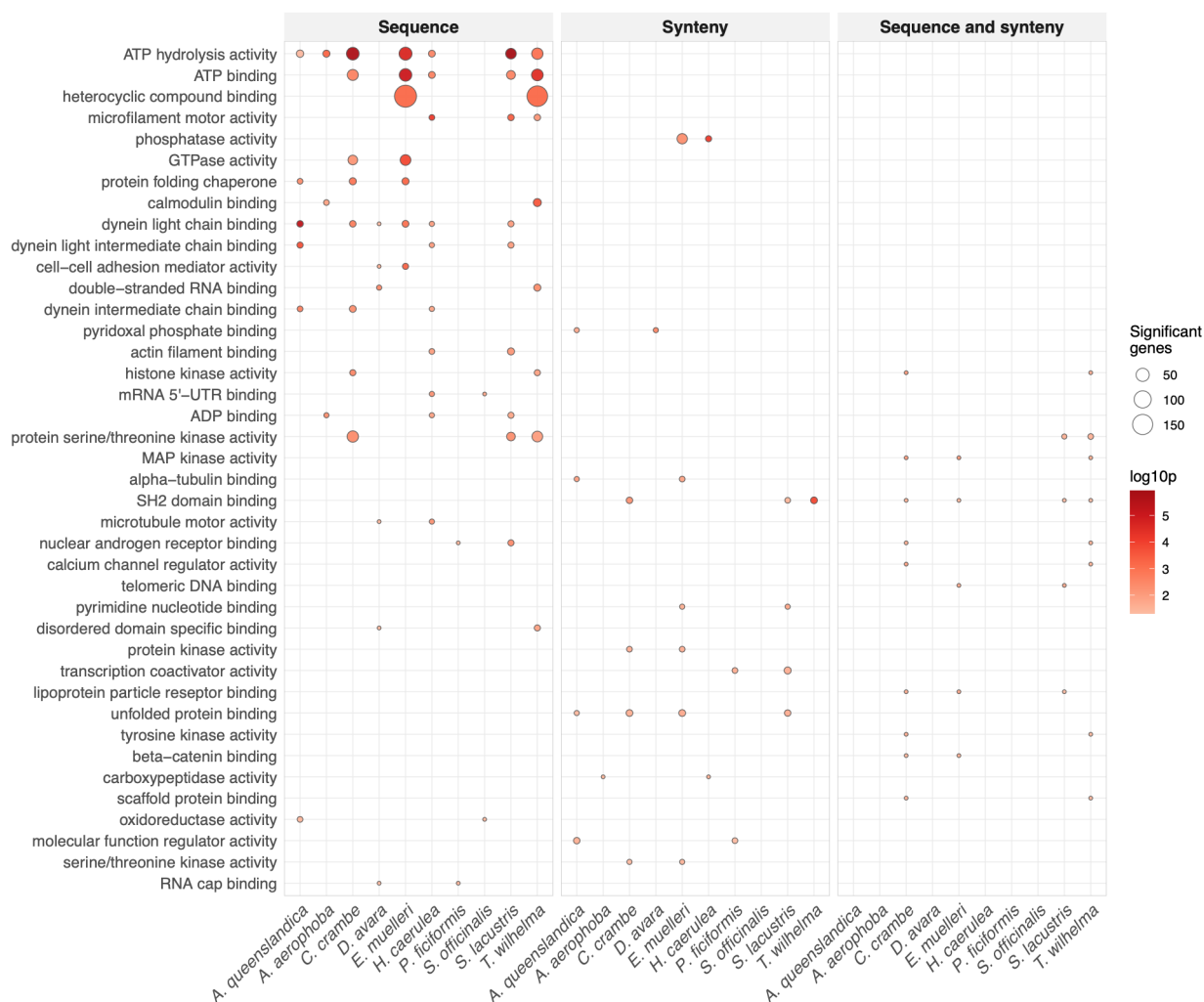


Figure 72. Molecular Function GO enrichment associated with conserved lncRNA loci across the analysed sponge species.

4.5.8. Long Non-coding RNA Expression Levels

The expression of lncRNA genes and BRAKER3-derived PCGs was compared across selected RNA-seq libraries. Only expressed genes (TPM > 0) were included in the analysis. For species represented by more than five libraries, a random subset of five libraries was selected for analysis. Across all analysed libraries, expressed lncRNA genes exhibited significantly lower expression levels than expressed PCGs ($p < 0.001$) (Figure 73). Although expression distributions partially overlapped, lncRNA genes were consistently skewed towards lower TPM values.

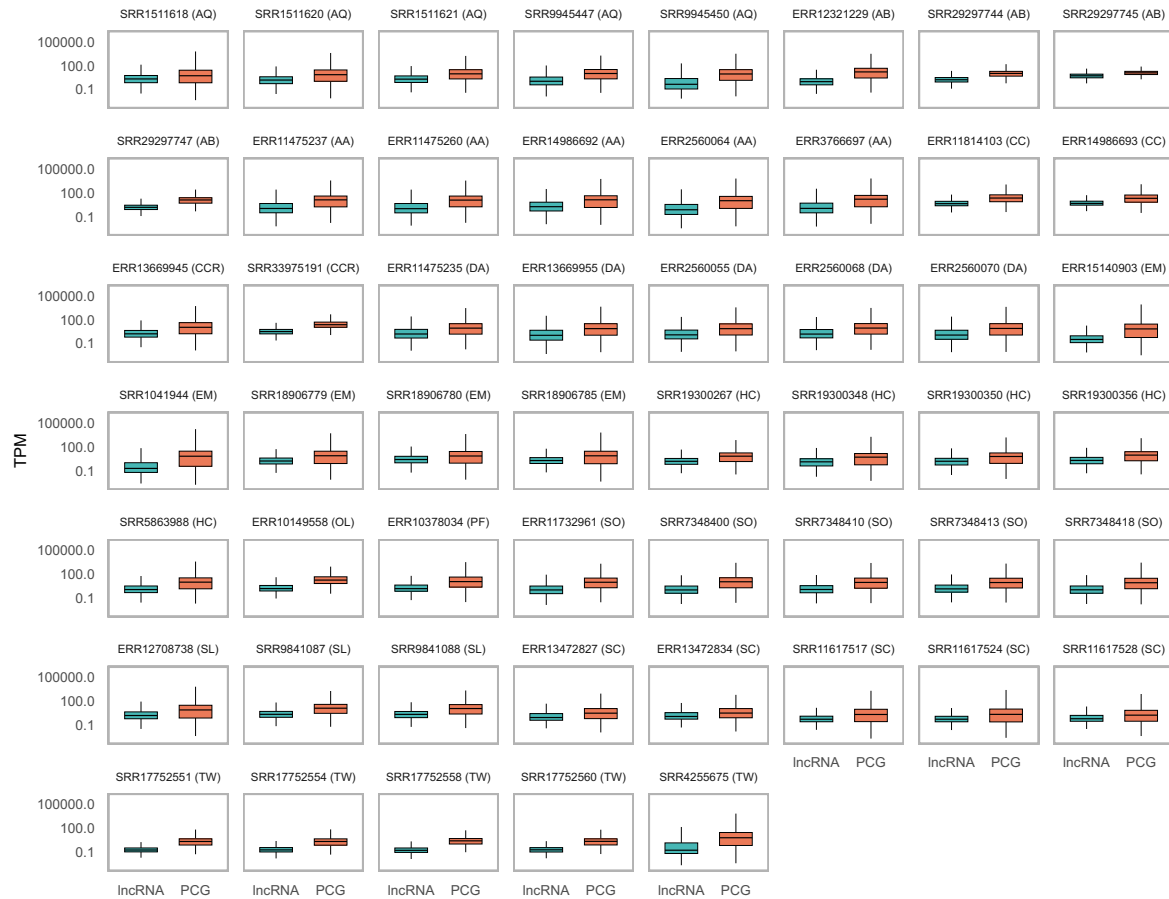


Figure 73. Expression comparison between expressed lncRNA genes and BRAKER3-derived PCGs across selected RNA-seq libraries. AQ = *A. queenslandica*, AB = *A. beatrix*, AA = *A. aerophoba*, CC = *C. candelabrum*, CCR = *C. crambe*, DA = *D. avara*, EM = *E. muelleri*, HC = *H. caerulea*, OL = *O. lobularis*, PF = *P. ficiformis*, SO = *S. officinalis*, SL = *S. lacustris*, SC = *S. ciliatum*, TW = *T. wilhelma*, TPM = transcripts per million, lncRNA = long non-coding RNA, PCG = protein-coding gene.

Expression patterns were additionally compared among different positional classes of lncRNAs, including intergenic, antisense, intronic, overlapping, and bidirectional classes (Figure 74). Overall, expression differences between positional categories were relatively modest across the analysed RNA-seq libraries. However, intergenic and bidirectional lncRNA genes frequently displayed slightly broader expression distributions and higher upper TPM ranges compared to other classes, suggesting the presence of a subset of more highly expressed genes within these categories. In contrast, antisense and intronic lncRNA genes generally exhibited narrower expression ranges and lower median expression values in several datasets. Despite these trends, substantial overlap in expression distributions was observed among all positional classes.



Figure 74. Expression comparison between lncRNA classes across selected RNA-seq libraries. Only expressed lncRNAs (TPM > 0) are included. AQ = *A. queenslandica*, AB = *A. beatrix*, AA = *A. aerophoba*, CC = *C. candelabrum*, CCR = *C. crambe*, DA = *D. avara*, EM = *E. muelleri*, HC = *H. caerulea*, OL = *O. lobularis*, PF = *P. ficiformis*, SO = *S. officinalis*, SL = *S. lacustris*, SC = *S. ciliatum*, TW = *T. wilhelma*, TPM = transcripts per million.

Finally, the proportion of RNA-seq libraries in which individual lncRNA genes were expressed was analysed to assess the extent of sample-specific expression (Figure 75). For each species, the proportion of libraries in which a given gene was detected (TPM > 0) was calculated. Across all species, a substantial fraction of lncRNA genes were expressed in a small subset of samples. In particular, many lncRNAs genes were expressed in fewer than 20% of libraries, whereas relatively few were expressed across all analysed samples. Species represented by larger RNA-seq datasets, including *A. aerophoba*, *D. avara*, *E. muelleri*, and *T. wilhelma*, exhibited especially pronounced enrichment of low-frequency expression categories. Conversely, only a minority of lncRNA genes showed constitutive expression across most or all samples.

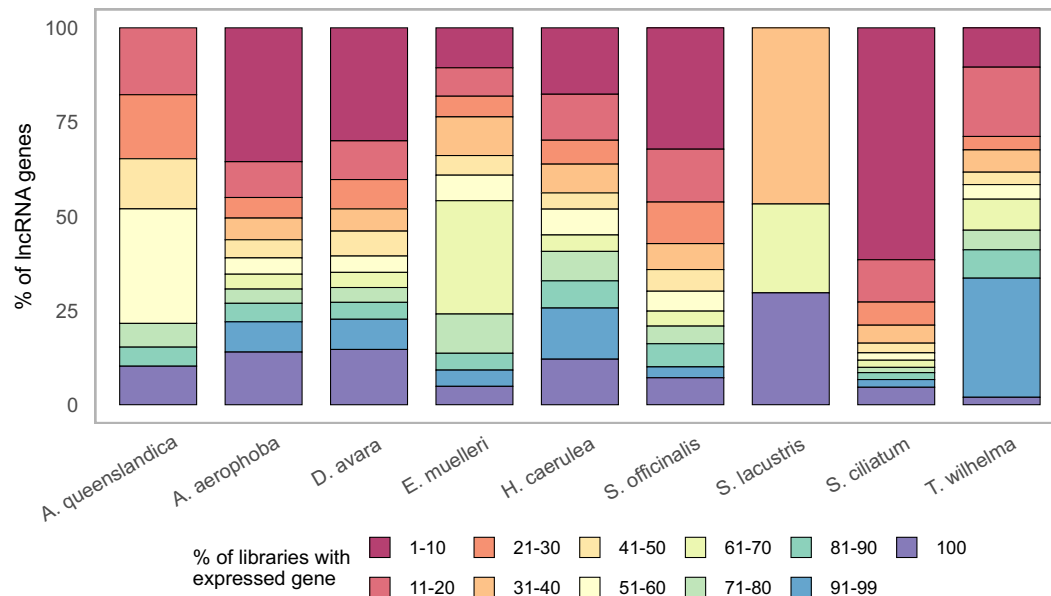


Figure 75. Proportion of RNA-seq libraries in which individual lncRNA genes were expressed (TPM > 0) across the analysed sponge species. lncRNA = long non-coding RNA.

4.5.9. Case Study: Differential Expression of Long Non-Coding RNAs in Response to Microbial-Associated Molecular Patterns

To investigate whether lncRNA expression responds to microbial immune stimulation, publicly available RNA-seq datasets generated by Pita et al. (2018) were reanalysed. In this study, two marine demosponges were exposed to MAMPs, and transcriptional responses were quantified at multiple time points following exposure.

The dataset comprised 35 RNA-seq libraries from *A. aerophoba* and *D. avara*, representing a subset of the transcriptomic datasets used for lncRNA characterization in this dissertation. For each species, individuals were exposed either to MAMPs or a control treatment and sampled at 1, 3, and 5 hours post-exposure, with three biological replicates per treatment and time-point combination. The only exception was the *D. avara* MAMP treatment at 1 h post-exposure, for which only two biological replicates were available. Differential expression analyses were performed separately for each species to assess changes in lncRNA gene expression following MAMP stimulation.

First, PCA of variance-stabilised lncRNA expression profiles revealed clear separation of samples according to treatment in both species (Figure 76). In *A. aerophoba*, separation between control and MAMP-treated samples was primarily captured along PC1, which explained 30% of the total variance, whereas temporal variation was reflected along PC2 (23% variance explained). The greatest divergence from control samples was observed for the 1 h MAMP treatment, while 3 h and 5 h samples occupied intermediate positions. A similar pattern was observed in *D. avara*, where PC1 and PC2 explained 31% and 20% of the gene expression variance, respectively. Control and MAMP-treated samples formed distinct

clusters, with additional structuring according to sampling time within treatment groups. Although several control samples in *D. avara* exhibited greater dispersion along PC2, no obvious outliers were detected that would justify exclusion from downstream analyses.

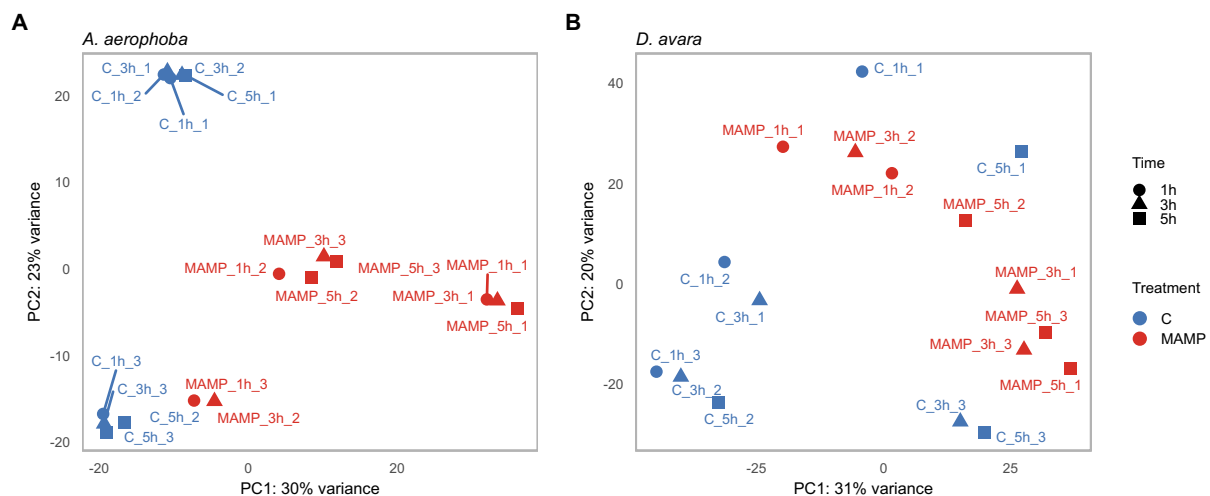


Figure 76. Principal component analysis of lncRNA gene expression profiles following MAMP exposure in *A. aerophoba* (A) and *D. avara* (B). Samples are coloured according to treatment (control or MAMP) and shaped according to sampling time (1, 3, or 5 h post-exposure). Numbers in sample labels indicate biological replicates. PC = principal component, C = control, MAMP = microbial-associated molecular patterns.

Differential expression analysis was performed separately for each species by comparing MAMP-treated and control samples at 1, 3, and 5 h post-exposure. Significant expression changes were detected in both lncRNA genes and PCGs, although the magnitude and direction of the transcriptional response varied between species and time points (Figure 77). In *A. aerophoba*, DEGs were predominantly upregulated throughout the experiment. At 1 h post-exposure, 91 PCGs and 51 lncRNA genes were upregulated, compared with 53 PCGs and 40 lncRNA genes that were downregulated. The number of DEGs increased at later time points, reaching 208 PCGs (133 upregulated, 75 downregulated) and 114 lncRNA genes (68 upregulated, 46 downregulated) at 3 h. The strongest transcriptional response was observed at 5 h, when 241 PCGs (159 upregulated, 82 downregulated) and 110 lncRNA genes (69 upregulated, 41 downregulated) were differentially expressed. Consistent with this predominance of upregulation, several DEGs exhibited large positive log₂ fold changes.

In contrast, *D. avara* exhibited a more balanced transcriptional response. At 1 h post-exposure, 63 PCGs and 49 lncRNA genes were differentially expressed, with upregulated genes slightly predominating in both categories. By 3 h, the number of DEGs had increased to 102 PCGs and 97 lncRNA genes, again with a modest excess of upregulated genes. At 5 h, however, lncRNA expression shifted towards repression, with 70 downregulated and 41 upregulated lncRNA genes identified. In contrast, PCGs retained a relatively balanced pattern, with 59 downregulated and 54 upregulated genes. Volcano plots additionally revealed several

lncRNAs exhibiting exceptionally large expression changes, including genes with $\log_2$ fold changes exceeding 20.

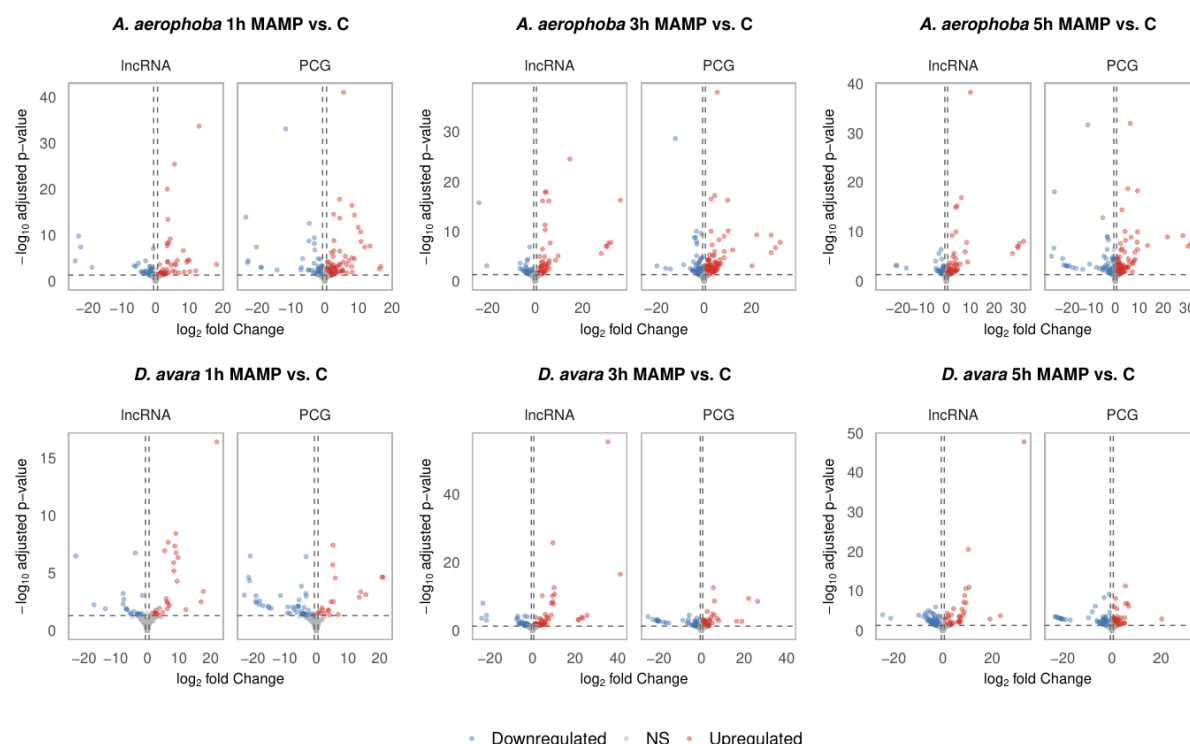


Figure 77. Differential expression of lncRNA genes and PCGs following MAMP exposure in *A. aerophoba* and *D. avara*. Each point represents a single gene. MAMP = Microbial-associated molecular patterns, C = control, lncRNA = long non-coding RNA, PCG = protein-coding gene, NS = non-significant.

To visualise the expression patterns of DEGs between MAMP-treated and control samples, variance-stabilised expression values of all differentially expressed lncRNA genes and PCGs were subjected to hierarchical clustering (Figure 78). In both *A. aerophoba* and *D. avara*, DEGs seemed to form distinct expression modules following MAMP exposure. Biological replicates generally clustered together within treatment groups. Notably, lncRNA genes and PCGs were interspersed throughout the expression clusters rather than forming separate groups. In *A. aerophoba*, the DEG set exhibited a pronounced division between genes upregulated and downregulated following MAMP exposure. A similar pattern was observed in *D. avara*, although the separation between induced and repressed expression modules was less distinct.

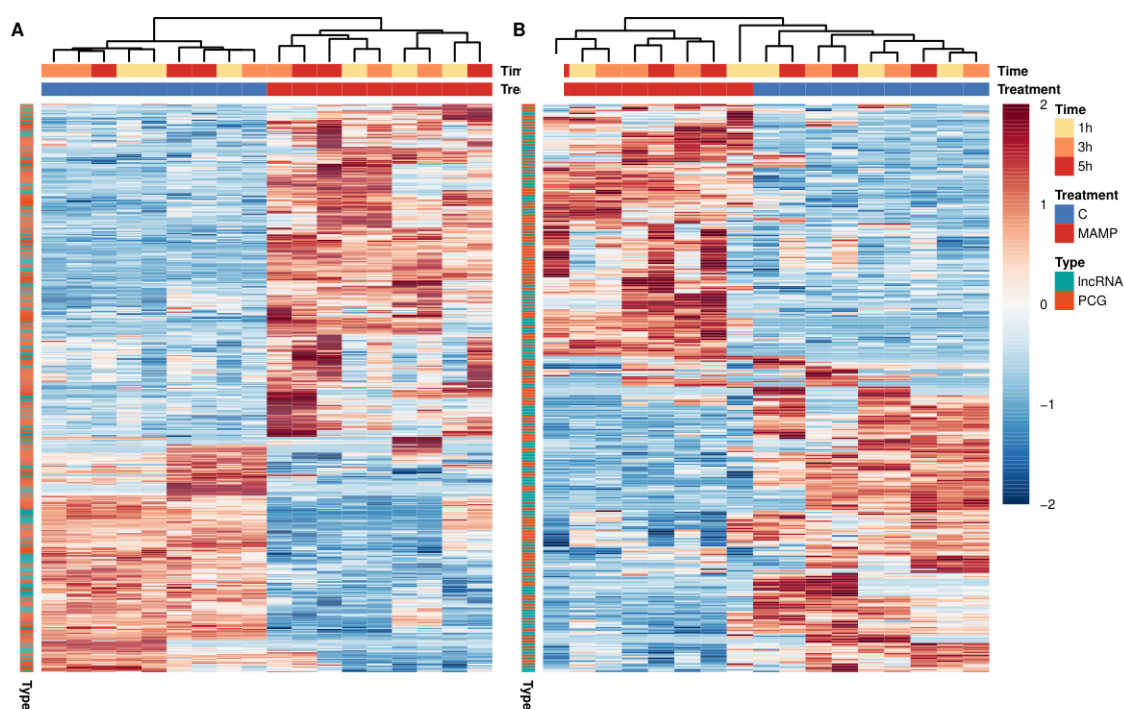


Figure 78. Hierarchical clustering of differentially expressed lncRNA genes and PCGs following MAMP exposure in *A. aerophoba* (A) and *D. avara* (B). Expression values are scaled by gene (row z-scores). C = control, MAMP = microbial-associated molecular patterns, lncRNA = long non-coding RNA, PCG = protein-coding genes.

To investigate coordinated transcriptional responses and identify lncRNA genes potentially co-regulated with PCGs, WGCNA was performed separately for each species. Co-expression networks were constructed using variance-stabilised expression values of all genes passing the low-count filter, retaining the top 75% most variable genes. Six co-expression modules per species were selected for downstream analysis based on significant correlation with MAMP treatment and balanced lncRNA and PCG representation.

Module eigengene profiles confirmed that the selected modules captured distinct transcriptional responses to MAMP stimulation (Figure 79). In *A. aerophoba*, the blue and greenyellow modules exhibited progressively increasing expression following MAMP exposure, reaching their highest levels at 5 h post-exposure. In contrast, the magenta and red modules were consistently downregulated in MAMP-treated samples across all time points. The tan module displayed a more transient pattern, showing reduced expression at 1–3 h followed by partial recovery at 5 h.

In *D. avara*, the greenyellow module showed the earliest and strongest induction, with elevated expression already evident at 1 h post-exposure and maintained throughout the experiment. Conversely, the magenta and red modules were consistently repressed in response to MAMP treatment. The yellow module exhibited a pronounced temporal trend,

with expression increasing progressively from 1 h to 5 h, whereas the black and lightgreen modules displayed more moderate treatment-associated changes.

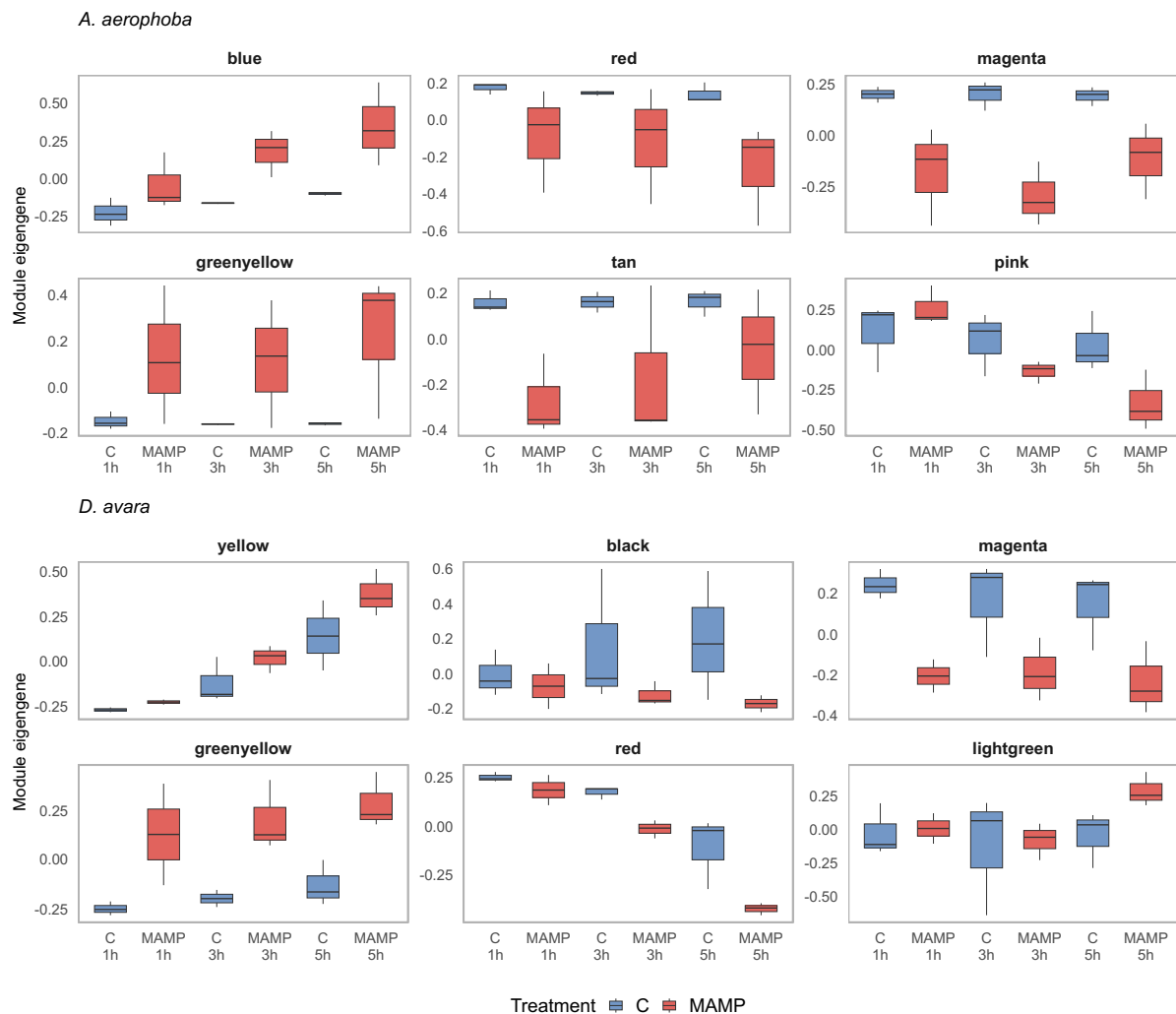


Figure 79. Module eigengene expression profiles across MAMP treatment conditions. Results are shown for the six selected MAMP-responsive co-expression modules in *A. aerophoba* (upper panels) and *D. avara* (lower panels), Module names correspond to WGCNA colour assignments. C = control, MAMP = microbial-associated molecular patterns.

To characterise the biological processes represented within each module, GO enrichment analysis was performed on the PCGs assigned to each selected module. In *A. aerophoba* (Figure 80), the blue module was enriched for terms associated with MAPK cascade signalling, G protein-coupled receptor signalling, and positive regulation of angiogenesis and cell growth. The red module, which was strongly repressed following MAMP exposure, was enriched for cell cycle and DNA replication terms including mitotic nuclear division, sister chromatid segregation, and DNA-templated replication. The greenyellow module, which showed a similar upregulation pattern, was enriched for immune-specific terms including regulation of T-helper 1 type immune response and interleukin-1 production. The pink module was

enriched for terms related to cytoplasmic translation and cotranslational protein targeting to the endoplasmic reticulum. The tan and magenta modules showed no significant GO enrichment.

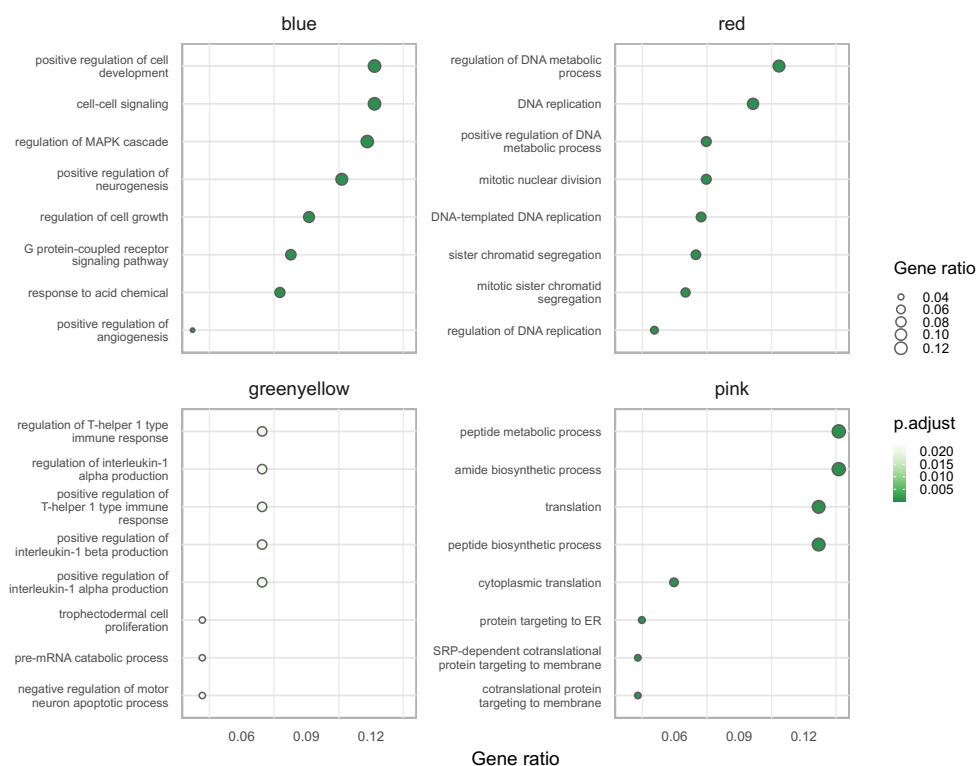


Figure 80. Biological Process GO enrichment of MAMP-responsive co-expression modules in *A. aerophoba*. Dot plots show the top eight enriched biological process GO terms for each selected co-expression module, ranked by statistical significance. Module names correspond to WGCNA colour assignments.

In *D. avara* (Figure 81), the yellow module was enriched for cell-cell signalling, neurone differentiation, and cholinergic synaptic transmission. The black module was enriched for NF- κ B signalling pathway terms, including regulation of I κ B phosphorylation and activation of NF- κ B-inducing kinase activity. The greenyellow module was enriched for terms related to bone mineralisation, regulation of ossification, and regulation of extracellular matrix disassembly. The lightgreen module was enriched for terms related to cytoplasmic translation and protein targeting to the endoplasmic reticulum. Magenta and tan modules did not exhibit any enriched GO terms

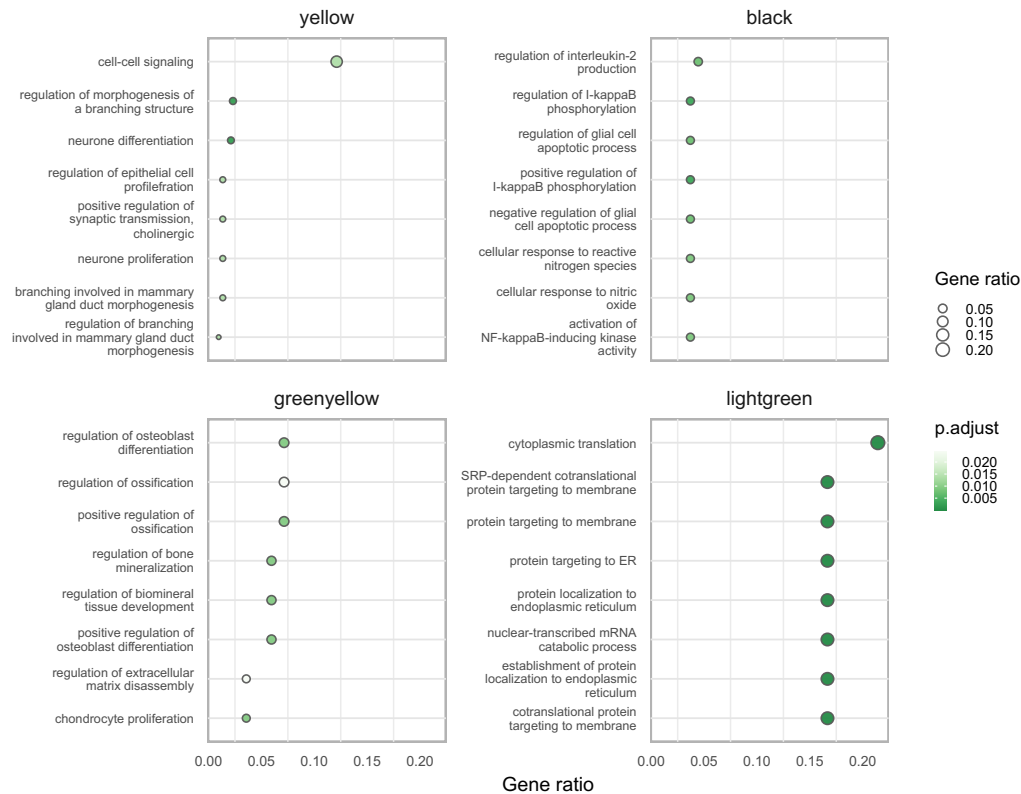


Figure 81. Biological Process GO enrichment of MAMP-responsive co-expression modules in *D. avara*. Dot plots show the top eight enriched biological process GO terms for each selected co-expression module, ranked by statistical significance. Module names correspond to WGCNA colour assignments.

To identify the most biologically relevant lncRNA candidates within each module, hub lncRNAs were defined as the top 10% most highly connected lncRNAs within their respective co-expression networks, based on the number of high-confidence co-expression partners. Hub lncRNAs were characterised by the number of network connections, module membership score, differential expression status, and genomic classification. Module membership scores were generally high, with the majority exceeding 0.85, indicating that hub nodes represent core module members rather than peripheral genes.

A total of 29 hub lncRNAs were identified across the six selected modules in *A. aerophoba* (Figure 82). These lncRNAs were predominantly intergenic, with antisense lncRNAs represented in the blue and tan modules. The most notable hub lncRNA was detected in the greenyellow module, which showed progressive MAMP-induced upregulation across all three timepoints with a maximum $\log_2$ fold change of 32.23 ($p < 0.001$) and a module membership score of 0.922. This lncRNA was classified as intronic. In the magenta module, characterised by strong MAMP-induced repression, one intergenic lncRNA was itself downregulated at 3 h post-exposure with a $\log_2$ fold change of 23.66 ($p < 0.001$) and a module membership of 0.907, representing the most significantly differentially expressed hub lncRNA identified in this

species. Two additional hub lncRNAs in the tan module were downregulated at 1 h post-exposure with $\log_2$ fold changes of 3.10 and 3.65, classified as antisense and intergenic, both with module membership scores above 0.89. The most highly connected hub in *A. aerophoba* was found in the red module, with 9 connections and a module membership of 0.927, though this lncRNA was not differentially expressed. Similarly, one lncRNA in the pink module showed the highest number of connections among pink module hubs, but was not a DEG.

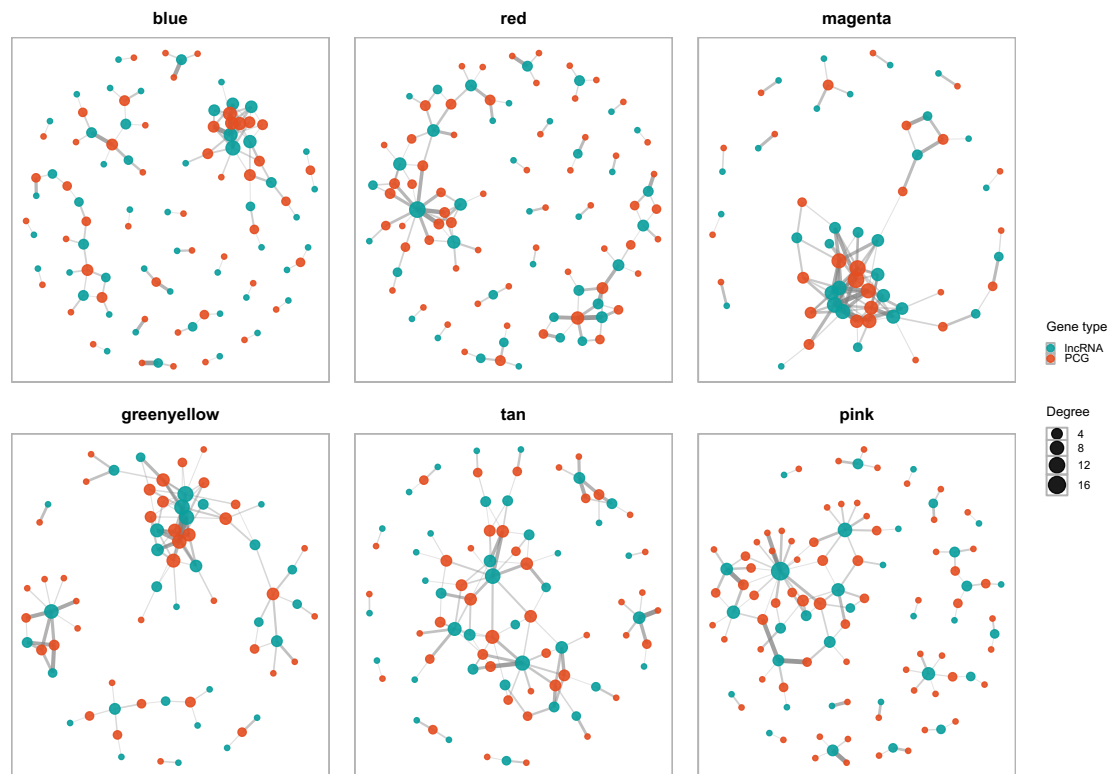


Figure 82. Network graphs for MAMP-responsive co-expression modules in *A. aerophoba*. Each panel represents a single co-expression module, with nodes coloured by gene type and sized proportionally to the number of co-expression partners within the network. Edges connect lncRNA–PCG pairs with Pearson correlation $r > 0.80$, with edge width and opacity scaled by correlation strength. Only the top 50 edges per module are shown for visual clarity. Module names correspond to WGCNA colour assignments. lncRNA = long non-coding RNA, PCG = protein-coding gene.

A total of 25 hub lncRNAs were identified in *D. avara* (Figure 83), with intergenic lncRNAs representing the predominant genomic class. The highest-confidence regulatory candidate identified in either species was a hub lncRNA in the greenyellow module, which represented the earliest and strongest MAMP-responsive module in *D. avara*. This intronic lncRNA was upregulated across all three timepoints with a maximum $\log_2$ fold change of 35.43 ($p < 0.001$) and a module membership score of 0.945. In the red module, one hub lncRNA showed modest downregulation at 5 h ($\log_2$ fold change = 1.78, $p = 0.002$) and was intergenic. The most highly connected hub lncRNA identified across both species was a hub lncRNA in the *D. avara*

magenta module, with 14 connected genes and a module membership of 0.878. Despite its exceptional connectivity, this lncRNA was not differentially expressed. In the black module, an intergenic lncRNA emerged as the dominant hub with eight connecting genes and a module membership of 0.848. This lncRNA was also not differentially expressed.

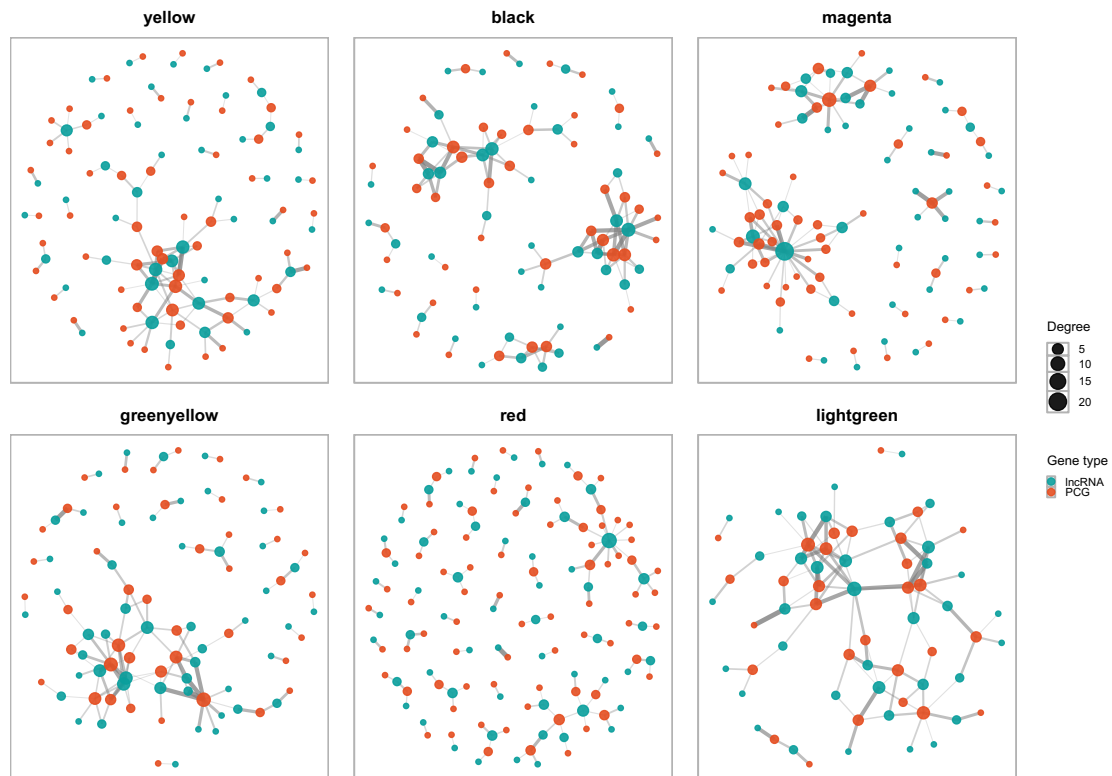


Figure 83. Co-expression network graphs for MAMP-responsive co-expression modules in *D. avara*. Each panel represents a single co-expression module, with nodes coloured by gene type and sized proportionally to the number of co-expression partners within the network. Edges connect lncRNA–PCG pairs with Pearson correlation $r > 0.80$, with edge width and opacity scaled by correlation strength. Only the top 50 edges per module are shown for visual clarity. Module names correspond to WGCNA colour assignments. lncRNA = long non-coding RNA, PCG = protein-coding gene.

5. DISCUSSION

This dissertation integrated methodological development with comparative transcriptomic and genomic analyses to investigate lncRNA repertoires in sponges. The work was framed by two central challenges. First, lncRNA identification in non-model organisms requires high-confidence transcriptome assemblies (Mattick et al., 2023; Ramakrishnaiah et al., 2020), yet existing transcriptome quality assessment tools are limited by insufficient interpretability, reduced accuracy under certain assembly conditions, or strict dependence on reference annotations (Li et al., 2014; Smith-Unna et al., 2016; Bushmanova et al., 2016). Second, sponge genomes are shaped by their holobiont biology and frequently contain contaminating sequences originating from associated microorganisms, creating a major obstacle for reliable genome annotation and downstream lncRNA discovery (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021). Addressing these challenges required a workflow that was necessarily more complex than standard model-organism pipelines. However, this complexity was essential considering that the analysed species differed substantially in genome quality and RNA-seq library availability, while representing a deeply divergent evolutionary radiation in which lncRNAs are expected to be weakly conserved and difficult to annotate.

Three computational tools were developed, evaluated, and applied within the scope of this dissertation. Two of these tools constitute the CATS framework. CATS-rf is a reference-free transcriptome quality assessment method based on RNA-seq read support, whereas CATS-rb is a reference-based method that evaluates transcriptome structural accuracy and completeness through transcript-to-genome alignment without requiring genome annotation. The third tool is an integrated genome decontamination pipeline that combines multiple lines of evidence to identify contaminant scaffolds. Together, these methods provided the quality-control framework required for high-confidence lncRNA characterization across the analysed sponge species. The resulting dataset represents the most comprehensive comparative analysis of lncRNAs in Porifera to date, extending previous studies that focused on individual species such as *A. queenslandica* (Gaiti et al., 2015), *S. ciliatum* (Fortunato et al., 2014), and *E. subterraneus* (Bodulić, 2020).

5.1. CATS Enables Comprehensive Transcriptome Assembly Evaluation in Non-model Organisms

Transcriptome assembly quality directly determines the reliability of downstream transcript discovery (Li et al., 2014; Smith-Unna et al., 2016). Consequently, robust assembly evaluation is essential for accurate lncRNA identification. To meet this need, CATS was developed by integrating two complementary transcriptome quality-assessment strategies. (Bodulić & Vlahoviček, 2026).

CATS-rf demonstrated superior performance compared with existing reference-free transcriptome assessment tools across a broad range of simulated libraries. Although the controlled simulations did not capture realistic intralibrary coverage variability, they provided

a framework for isolating the effects of diverse conditions on CATS scoring. When compared to existing quality assessment tools, CATS-rf offered significantly more accurate quality metrics at non-minimum sequencing depth across all analysed species and assemblers, while remaining highly competitive at the lowest coverage level (1-4×). The realistic simulations extended these findings to more biologically plausible datasets. Under these conditions, CATS-rf maintained strong associations with transcript quality, outperforming RSEM-EVAL and TransRate across the vast majority of the analysed species, sequencing depth, and assemblers.

CATS-rf also performed significantly better than existing reference-free assessment tools on public RNA-seq datasets, although its quality estimates showed weaker correlations with transcript-level assembly accuracy compared to simulated benchmarks. This reduction is expected, as benchmarking on real datasets is inherently limited by the absence of complete ground-truth transcriptomes, with reference transcriptomes biased toward well-annotated transcripts. Consequently, assembled transcripts absent from the reference may receive low F-scores despite representing bona fide transcripts. To reduce this source of bias, benchmarking was restricted to transcripts represented in the reference transcriptome, as determined by alignment-based criteria. However, this filtering may also have excluded poorly assembled transcripts that failed to align to the reference. As a result, the evaluation set was likely enriched for higher-quality transcripts. Therefore, the public-dataset benchmark should be interpreted alongside the simulation results, which provide a more controlled assessment of CATS performance.

Reference-free approaches are generally limited in their ability to estimate absolute transcript quality across different libraries (Li et al., 2014; Smith-Unna et al., 2016). This limitation is particularly evident for RSEM-EVAL, whose transcript-level metrics primarily reflect each transcript's contribution to the overall assembly score rather than its intrinsic accuracy (Li et al., 2014). In contrast, CATS-rf and TransRate were both designed to estimate transcript-level quality directly from read-mapping evidence (Smith-Unna et al., 2016). Notably, CATS-rf consistently ranked assemblies according to objective transcript accuracy across diverse simulated conditions and outperformed the transcript-level scores produced by both RSEM-EVAL and TransRate. These results suggest that the read-mapping-based error estimates implemented in CATS-rf provide a reliable proxy for absolute transcript accuracy.

Although CATS-rf and TransRate share the general principle of using read mapping and transcript-level scoring to quantify assembly errors, they differ substantially in their underlying methodology. Key distinctions include their approaches to read mapping, assignment of multimapped reads, definition of score components, and calculation of the overall assembly score (Smith-Unna et al., 2016). The superior performance of CATS-rf indicates that the specific definition of score components is critical for distinguishing genuine assembly errors from mapping inconsistencies caused by sequencing noise or biological variation. The behaviour of CATS-rf score components was further supported by mutation introduction experiments. In transcripts with sufficient read coverage, individual components

responded as expected to different classes of introduced errors. However, this sensitivity may be constrained at minimum sequencing depth, where limited read evidence can obscure the mapping patterns required to identify specific misassembly events. This likely contributed to slightly lower accuracy for low-coverage transcripts in the simulation experiments.

Although CATS-rf score components cannot be considered exclusively specific to a single error type, the low intercorrelations among components across diverse assembly conditions suggest that they capture distinct aspects of assembly quality. Similarly, the separation of mutation types in PCA indicates that different classes of misassembly generate characteristic CATS scoring profiles. Together, these findings support that CATS-rf components provide informative, complementary signals rather than redundant measures of transcript quality.

While CATS-rf score components penalize unsupported fusion sites, CATS-rf does not directly detect chimerism. This limitation is likely minor considering that several studies report a low incidence of chimeric transcripts produced by modern assemblers (Honaas et al., 2016; Wang & Gribskov, 2017; Cabau et al., 2017). In contrast, TransRate computes a chimerism score based on the assumption of uniform coverage (Smith-Unna et al., 2016), which may be unrealistic given known sources of coverage heterogeneity, including fragmentation and GC-content bias (Patterson et al., 2019; Li et al., 2012; Lahens et al., 2014). Meanwhile, the reference-based methodology implemented in CATS-rb demonstrated strong performance in identifying chimerism. However, the obtained results should be interpreted in the context of rare-event classification.

CATS-rb provides a second major contribution of the CATS framework by estimating transcriptome assembly completeness without strictly requiring a reference transcriptome or genome annotation. This feature is particularly important for non-model organisms and lncRNA studies, where annotations are often incomplete or unavailable. Unlike annotation-dependent approaches, CATS-rb defines non-redundant transcript and exon element sets directly from transcript-to-genome alignments and evaluates their representation across assemblies. The strong correlations between relative and annotation-based CATS-rb scores in controlled and realistic datasets indicate that this relative scoring strategy provides a reliable estimate of transcriptome completeness. Reference transcriptomes consistently received near-maximum scores, while mutated assemblies showed progressively reduced completeness, demonstrating that CATS-rb is sensitive to assembly errors without substantial score inflation. These findings suggest that CATS-rb is valuable in several analytical contexts, including the comparison of assemblies generated from the same library, the evaluation of transcriptomic content across libraries, and the assessment of absolute completeness when a reference transcriptome or genomic annotation is available.

CATS-rb scores were also strongly associated with CATS-rf assembly scores, indicating broad consistency between the two CATS modules. However, these metrics differ in their interpretation. CATS-rf scores primarily assess the accuracy of the assembled transcripts,

whereas CATS-rb captures the completeness of reconstructed transcriptomic content. This distinction was most apparent at the lowest coverage levels, where CATS-rb scores slightly diverged from accuracy-based metrics. Therefore, CATS-rf and CATS-rb results should be interpreted jointly, as their combined use provides a more comprehensive representation of transcriptome assembly quality.

The performance of both CATS modules is intrinsically linked to their respective alignment strategies. CATS-rf uses Bowtie2 with optimized parameters designed to balance mapping stringency and error detection, whereas CATS-rb uses Spaln for transcript-to-genome alignment with species-specific parameters that can be adjusted by the user. Although these aligners are not considered universally optimal, both are regarded as highly accurate, and the selected parameters were optimized for transcriptome assembly evaluation (Langmead & Salzberg, 2012; Gotoh, 2024).

Benchmarking analyses revealed several consistent biological and technical determinants of transcriptome assembly quality. More complex transcriptomes, particularly those of vertebrates, consistently yielded lower assembly scores, reflecting the inherent challenges associated with reconstructing isoform-rich transcriptomes containing extensive paralogy, alternative splicing, and repetitive sequence content (Raghavan et al., 2022; Jackson et al., 2024). Across both CATS modules, rnaSPAdes and Trinity generally outperformed IDBA-tran and SOAPdenovo-Trans, in agreement with previous benchmarking studies (Hölzer et al., 2019; Wang & Gribskov, 2017). However, these comparisons should be interpreted cautiously, as only default assembler parameters was evaluated.

Sequencing depth exerted a strong positive effect on assembly quality at low coverage levels, although improvements progressively diminished as coverage increased, particularly in libraries with low sequencing error rates. Similarly, coverage increases in realistic simulations yielded progressively smaller gains in assembly quality, considering that most reads originated from strongly expressed transcripts. These results are consistent with previous studies demonstrating diminishing returns of increased sequencing depth for transcriptome reconstruction (Patterson et al., 2019; Tarazona et al., 2011). Collectively, these findings suggest that, beyond moderate coverage levels, improvements in transcriptome completeness may be more effectively achieved through broader biological sampling, including the incorporation of diverse tissues, developmental stages, and experimental conditions.

Overall, CATS offers a comprehensive framework for transcriptome assembly evaluation by combining the strengths of reference-free and reference-based approaches. By providing interpretable and reliable metrics, CATS can support quality control and assembly selection in a broad range of transcriptomic studies, including applications involving poorly annotated organisms and non-coding RNA discovery.

5.2. *Sponge Transcriptome Assembly Required Combined De Novo, Genome-guided, and Clustering Strategies*

Transcriptome assembly in non-model organisms benefits from the integration of multiple reconstruction strategies (Raghavan et al., 2022; Hölzer & Marz, 2019). In this dissertation, an integrated assembly strategy was required as the analysed sponge species differed markedly in genome contiguity, contamination levels, transcriptome complexity, and available RNA-seq depth. Consequently, no single assembler or parameter configuration was expected to perform optimally across all species. Instead, the combination of de novo and genome-guided assembly outputs provided a more flexible framework for recovering transcript diversity.

Transcriptome reconstruction was therefore based on three complementary assembly approaches: rnaSPAdes de novo assembly, Trinity de novo assembly, and Trinity genome-guided assembly. rnaSPAdes and Trinity were selected as they represent well-established assemblers for short-read transcriptome reconstruction. In addition, both tools support strand-specific assembly, which was essential for accurate lncRNA classification. By contrast, IDBA-tran and SOAPdenovo-Trans were excluded from the sponge transcriptome workflow due to their poor performance in the CATS benchmark and their limited suitability for stranded RNA-seq assembly. StringTie, a widely used reference-guided assembler, was also evaluated as a potential component of the assembly workflow. However, it was not incorporated into the final pipeline due to its strong dependence on genome quality and existing genomic annotation. Although reference-guided assemblers are highly effective when high-quality genome assemblies and annotations are available, their output is inherently constrained by the reference data quality (Pertea et al., 2015). This limitation is particularly relevant in sponges, where genome assemblies differ substantially in contiguity and accuracy. Trinity genome-guided assembly was therefore selected as a more flexible alternative, in which the reference genome is used solely to guide read alignment and transcript reconstruction, while allowing transcript structures to be inferred directly from the RNA-seq data (Grabherr et al., 2011).

The de novo assembly results revealed a consistent trade-off between accuracy and completeness. Although rnaSPAdes assemblies generally showed higher transcript accuracy, their quality varied among species, indicating that rnaSPAdes performance was sensitive to species-specific differences in transcriptome complexity and library composition. Trinity de novo assemblies were generally more complete, but displayed lower CATS-rf scores and weaker contiguity. This suggests that Trinity recovered a broader set of transcript sequences, but at the cost of increased structural uncertainty. In contrast, rnaSPAdes appeared to be more conservative, producing more accurate transcript models but lower transcriptome completeness in several species. These complementary properties justified the integration of both assemblers in the final assembly workflow.

To improve transcriptome representation, the best-performing rnaSPAdes assembly, Trinity de novo assembly, and Trinity genome-guided assembly were merged for each species. The combined assemblies were subsequently subjected to redundancy reduction at several clustering stringencies. Redundancy reduction was necessary as the merged assemblies were generated from the same RNA-seq libraries and therefore were expected to contain substantial transcript overlap. Furthermore, transcriptome assemblies themselves frequently contain overlapping contigs, fragmented isoforms, and similar transcript variants. Together, these sources of redundancy can inflate lncRNA counts and bias downstream analyses. However, redundancy removal also introduced the possibility of collapsing closely related isoforms or paralogous transcripts. This risk was lowered by evaluating several clustering thresholds and selecting final assemblies primarily according to CATS-rb relative transcript score, while also favouring less redundant assemblies when completeness remained comparable.

Overall, the final optimised transcriptome assemblies represented a compromise among transcript accuracy, transcriptome completeness, and redundancy reduction. In several species, Trinity genome-guided assemblies were selected, indicating that genomic anchoring improved transcript reconstruction sufficiently to outweigh the expected completeness of clustered de novo assemblies. In most other species, combined assemblies were selected, demonstrating that the integration of de novo and genome-guided information provided the most complete representation of transcriptomic content. These findings support the central premise of the workflow: reliable lncRNA discovery in non-model organisms requires systematic quality assessment and species-specific assembly selection.

5.3. *Genome Decontamination Is Essential for Reliable Sponge lncRNA Analysis*

The rapid expansion of genomic resources for non-model metazoans has substantially advanced the study of early animal evolution, while introducing important challenges related to genome assembly quality and data integrity. These difficulties are particularly pronounced in sponges, which are holobionts that maintain close associations with bacteria, archaea, microbial eukaryotes, and viruses (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021). This biological complexity creates a major obstacle for genome assembly and annotation, as host-derived DNA cannot always be fully separated from symbiont-derived or environmental DNA during sequencing and assembly. In this dissertation, genome decontamination was therefore not treated as a peripheral preprocessing step, but as a prerequisite for reliable sponge lncRNA identification.

To address these challenges, the decontamination pipeline was developed combining three independent evidence types: sequence composition, protein-level taxonomy, and nucleotide-level signal (Bodulić & Vlahoviček, 2026). Validation using a spike-in dataset demonstrated strong pipeline performance, with high overall accuracy, precision, and recall, indicating that the integrated scoring system reliably distinguished sponge- and contaminant-derived

scaffolds under controlled conditions. The high precision was particularly important considering that false-positive removal of host scaffolds would compromise downstream gene annotation and lncRNA discovery. Comparisons of BUSCO scores between original and decontaminated sponge genomes further supported the specificity of the method, as metazoan genome completeness remained largely unchanged after decontamination. Together, these results indicate that the pipeline successfully removed non-target DNA while preserving the integrity of the host gene space.

However, the validation results should also be interpreted with caution. The high recall observed in the spike-in dataset may represent an optimistic estimate of real-world performance, considering that the source sequences included in the validation dataset were well represented in the databases utilized by the similarity-based components of the pipeline. In novel or poorly represented taxa, similarity-based signals may be weaker, and decontamination efficacy is therefore inherently constrained by the phylogenetic breadth of available sequence databases. This limitation is particularly relevant for non-model organisms such as sponges, where both host genomes and sponge-associated microbial lineages remain underrepresented. Under such conditions, the compositional score becomes particularly important considering that it provides a line of evidence that is independent of external database content.

The validation dataset was intentionally broad, containing fragments from multiple sponge genomes and diverse contaminant taxa. Although this design did not fully reproduce the structure of a typical single-species genome assembly, it provided a rigorous test of pipeline performance across a wide taxonomic range. In particular, the compositional component was challenged by the inclusion of multiple sponge genomes, which increased host-side sequence heterogeneity. The strong performance of the integrated score despite this increased compositional diversity indicates that the pipeline can tolerate substantial variation in sequence composition.

Concordance among scoring components differed substantially between the validation dataset and the sponge assemblies analysed for contamination. In the controlled spike-in dataset, a large proportion of contaminant sequences was identified by all three components. By contrast, in the analysed sponge genomes, most contaminant scaffolds were supported by only two components. This pattern underscores the necessity of the integrated scoring framework. Reliance on a single evidence type would increase the risk of false-positive classification. Conversely, requiring concordance among all three components would substantially decrease contaminant recovery. The two-out-of-three consensus criterion therefore provided an appropriate balance between precision and recall in underrepresented holobiont genomes.

Kingdom-level taxonomic aggregation represented an important design feature for the analysis of non-model metazoan genomes. Considering that sponge sequences are

underrepresented in public databases, sponge-derived genes may produce their strongest similarity matches to other metazoans rather than to closely related sponge sequences. Collapsing all metazoan hits into a single target category therefore reduced the risk of incorrectly classifying host-derived scaffolds as contaminants. However, this strategy also limited the resolution of taxonomy-based scoring components, as contaminants originating from the same kingdom could not be distinguished from host sequences. Consequently, the pipeline was not designed to detect intra-kingdom contamination, including human genetic material introduced during laboratory handling or sequences derived from other marine metazoans. This limitation was acceptable within the scope of this dissertation, as the most probable and biologically relevant contaminants in sponge assemblies were expected to originate from microbial symbionts rather than from other animals (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021).

The exclusion of scaffolds shorter than 1 kb represented another necessary compromise. Short scaffolds often lack sufficient k-mer diversity for robust compositional clustering and rarely contain enough coding sequence for confident taxonomic assignment (Astashyn et al., 2024; Bálint et al., 2024). Although some genuine host-derived sequences may be removed by this threshold, such short scaffolds are more likely to represent assembly artefacts or repeat fragments. Their exclusion therefore likely reduced noise in downstream annotation and lncRNA prediction.

The decontamination analysis revealed variable contamination across the sponge genomes examined in this dissertation. Although contaminant scaffolds in certain assemblies represented a substantial proportion of scaffold counts, they generally accounted for only a small fraction of total genome length. This pattern indicates that contamination was mostly associated with shorter scaffolds. Nevertheless, such scaffolds can still contain predicted genes or attract transcript mappings, and their removal was therefore important for ensuring the reliability of downstream analyses.

The taxonomic composition of detected contaminants reflected the biological complexity of the sponge holobiont. Contaminant scaffolds were predominantly assigned to bacterial lineages, consistent with the established importance of bacterial symbionts in sponge-associated microbial communities (Hentschel et al., 2012; Thomas et al., 2016; Robbins et al., 2021). Additional bacterial groups, as well as archaeal and algal sequences, were also identified in some assemblies. This variation likely reflects both ecological differences among sponge species and differences in the decontamination procedures applied in the original genome studies. The detection of archaeal and algal sequences is consistent with the known occurrence of ammonia-oxidising archaea and photosynthetic microbial eukaryotes in sponge holobionts (Pawlik & McMurray, 2020; Hentschel et al., 2012). These findings suggest that contamination in sponge genome assemblies often reflects the close biological association between the sponge host and its microbial community, rather than technical contamination introduced during laboratory processing.

No viral contaminants were identified in the analysed sponge genomes. Considering that the decontamination pipeline did not successfully recover viral sequences from the validation dataset, this absence most likely reflects methodological limitations rather than a true lack of viral signal. Viral scaffolds are often short and poorly represented in general nucleotide and protein databases, limiting their detection by the evidence types implemented in the current pipeline. Future versions of the decontamination framework would therefore benefit from the incorporation of dedicated viral detection approaches, including curated virus-specific reference databases and hallmark gene screening.

The presence of HGT poses an additional challenge for genome decontamination, as genuine host-integrated sequences of microbial origin may resemble contaminant scaffolds. Such regions can retain microbial-like compositional signatures despite being an integral part of the host genome (Pita et al., 2018; Hentschel et al., 2012; Kelly et al., 2022). The pipeline was not explicitly benchmarked against a curated set of confirmed HGT loci. However, the validation dataset included sponge chromosomal fragments that likely contained ancient HGT-derived regions, yet still yielded very high precision. This result suggests that the pipeline can effectively distinguish bona fide HGT-derived sequences from exogenous microbial contaminants.

Overall, these results demonstrate that genome contamination is present in sponge assemblies but can be effectively addressed using a consensus-based framework. By integrating compositional and taxonomy-based evidence, the pipeline achieved high accuracy while preserving host genome completeness. For the purposes of this dissertation, this step was essential as reliable lncRNA discovery depends on the accurate definition of host genomic space. Decontaminated assemblies therefore provided a more reliable foundation for gene prediction, transcriptome mapping, repeat annotation, and downstream characterization of sponge lncRNAs.

5.4. Repeat and Protein-coding Gene Annotation Revealed Substantial Genome Heterogeneity Among Sponges

Robust TE and PCG annotation provided an essential genomic framework for lncRNA identification and annotation. Because lncRNAs are defined not only by their transcript features but also by their relationship to neighbouring genomic elements, accurate contextual annotation was necessary for downstream analyses. Repeat annotations enabled the assessment of TE-derived contributions to lncRNA sequences, whereas PCG annotations provided the basis for PCT exclusion, lncRNA class assignment, synteny analysis, and functional interpretation. Applying a consistent annotation workflow across all 14 species was therefore important for reducing methodological variation in lncRNA characterization.

The repeat annotation strategy combined several state-of-the-art tools, thereby integrating structural, homology-based, and machine-learning classification approaches. This strategy was necessary considering that sponge repeats remain poorly represented in curated

databases, and reliance on a single repeat-detection method would likely underestimate repeat diversity. Even so, a large proportion of repetitive sequences in all analysed genomes remained unclassified, accounting for approximately 26–36% of the total repeat content. This substantial fraction of unknown repeats likely reflects the evolutionary divergence of sponge-specific TEs and the limited representation of poriferan repeat families in current reference databases.

Repeat content varied markedly among sponge genomes, ranging from approximately one-fifth to nearly three-quarters of total genome length. DNA transposons represented the dominant repeat class in all species, with LTR retrotransposons and LINEs also contributing substantially across several genomes. These findings are consistent with previous genomic studies indicating that repetitive DNA constitutes a major and highly variable component of sponge genomes. For instance, the published genome annotation of *A. queenslandica* contains a large repetitive fraction dominated by DNA transposons (Srivastava et al., 2010). Similarly, the public annotation of *E. muelleri* genome and the recently published genomes of *Tethya* species have revealed similar distributions of repeat landscapes and lineage-specific expansions of TE families (Kenny et al., 2020; Francis et al., 2017; Wörheide et al., 2024). A recent comprehensive study has further demonstrated considerable variation in repeat abundance among species, indicating that TEs have contributed substantially to genome size evolution and structural diversification across Porifera (Kuzman, 2020). The established variation in TE annotation among species may reflect genuine biological differences in TE activity and genome evolution, but may also be influenced by technical factors, including assembly quality and the ability to reconstruct repetitive regions.

Annotation of PCGs revealed substantial interspecific variation in gene number and genomic coverage. While some of this variation likely reflects genuine biological differences among sponge genomes, technical factors are also expected to contribute. Genome contiguity, repeat masking efficiency, RNA-seq depth, and the availability of homologous protein evidence can all influence gene prediction performance and the resulting BRAKER3 annotations (Gabriel et al., 2024). The use of a standardized BRAKER3 workflow across all species minimized methodological bias and improved the comparability of annotations, although differences arising from variation in genome assembly quality and supporting evidence could not be entirely eliminated.

Importantly, PCG annotations generated in this study were broadly consistent with published sponge genome annotations, which also report substantial variation in predicted gene content among species such as *A. queenslandica* (Srivastava et al., 2010), *E. muelleri* (Kenny et al., 2020), and *T. wilhelma* (Francis et al., 2017; Wörheide et al., 2024). However, direct comparison with previous studies should be made cautiously, as published annotations were produced using different genome assemblies, evidence sets, repeat-masking procedures, and gene-prediction workflows. Species-specific genome projects often incorporate tailored annotation refinement, whereas the present study prioritized a uniform strategy across all

analysed species. The use of decontaminated and soft-masked genome assemblies may also have contributed to differences from previously published annotations. Removal of contaminant scaffolds reduces the risk of inflated gene counts caused by microbial sequences, while repeat masking limits false predictions within repetitive regions. Consequently, the annotations generated here may be conservative for some species, but they provide a consistent protein-coding reference framework for lncRNA classification and PCT filtering. This consistency was more important for the comparative aims of the dissertation than maximising the absolute number of bona fide PCGs in each species.

5.5. Sponge Long Non-coding RNA Repertoires are Extensive but Strongly Influenced by RNA-seq Sampling

The lncRNA identification pipeline recovered extensive repertoires of candidate lncRNAs across all analysed sponge species. Final lncRNA counts varied substantially among species, ranging from several thousand to tens of thousands of transcripts. However, these differences should not be interpreted solely as biological variation in lncRNA repertoire size. They are strongly influenced by RNA-seq sampling, including the sampled tissue, developmental stage, experimental condition, the number of available libraries, sequencing depth, and library type. Species represented by numerous RNA-seq libraries sampled across diverse biological contexts are expected to yield larger lncRNA catalogues than species represented by only one or two libraries, particularly because lncRNAs are often lowly expressed and sample-specific (Statello et al., 2021; Mattick et al., 2023).

Several species also exhibited substantial reductions in lncRNA candidate numbers following the removal of transcripts that could not be mapped to the corresponding decontaminated genome assembly. This observation likely reflects the presence of transcripts originating from sponge-associated microorganisms, which are common components of sponge RNA-seq datasets. This finding highlights the importance of genome-based filtering for accurate lncRNA identification in holobiont systems.

The stringent filtering strategy was designed to minimise low-quality transcripts and false-positive lncRNA predictions. However, several classes of bona fide lncRNAs were likely excluded during this process, including mono-exonic transcripts, lowly expressed lncRNAs, transcripts containing subregions similar to structural RNAs, lncRNAs originating from pseudogenes, and sense lncRNAs. This trade-off was considered acceptable considering that the primary objective was to generate a high-confidence lncRNA dataset with minimal contamination from coding transcripts. Accordingly, the workflow incorporated multiple safeguards, including coding-potential prediction, searches against protein databases, and removal of transcripts overlapping annotated PCGs. This conservative approach increased confidence that the final lncRNA catalogue was not substantially contaminated by unannotated coding transcripts.

This work substantially expands the known sponge lncRNA repertoire. Previous analyses identified 2935 lncRNAs in *A. queenslandica* (Gaiti et al., 2015), approximately 600 intergenic lncRNAs in *S. ciliatum* (Fortunato et al., 2014), and 4913 lncRNAs in *E. subterraneus* (Bodulić, 2020). By contrast, the present analysis detected substantially larger candidate lncRNA sets in *A. queenslandica*, *S. ciliatum*, and 12 additional sponge species. This increase is expected given that the present workflow integrated all available stranded RNA-seq libraries that met quality criteria, applied combined transcriptome assembly strategies, and did not restrict candidate identification to intergenic lncRNAs. Direct comparison with previously reported lncRNAs from *A. queenslandica* and *S. ciliatum* was not possible as these lncRNA sets are not publicly available. These results demonstrate that broader RNA-seq sampling and improved assembly and annotation workflows recover a substantially larger and more diverse set of sponge lncRNAs.

5.6. *Sponge Long Non-coding RNAs Share Major Structural Features with lncRNAs from Other Animals*

Despite variation in RNA-seq sampling, genome quality, and transcriptome assembly completeness, several consistent patterns were observed across sponge species. First, lncRNA transcripts and genes were shorter than both assembly-derived and BRAKER3-derived protein-coding counterparts in all species. Furthermore, lncRNAs contained fewer exons and generally exhibited lower isoform numbers per gene. These structural features are consistent with lncRNA studies in bilaterians and other eukaryotes (Statello et al., 2021; Mattick et al., 2023; Kung et al., 2013). The presence of these characteristics in sponges suggests that major structural properties of animal lncRNAs were already established early in metazoan evolution.

The comparison between assembly-derived and BRAKER3-derived PCTs was important for interpreting several observed patterns. Assembly-derived PCTs were generated from the same transcriptome assemblies as lncRNAs and therefore provided the most appropriate control for assembly-related effects. In contrast, BRAKER3-derived PCTs provided an annotation-based coding reference, with gene models primarily encompassing protein-coding regions and frequently lacking complete untranslated regions. This methodological distinction likely accounts for the greater lengths observed in assembly-derived PCTs relative to their BRAKER3-derived counterparts. Importantly, the consistent pattern of shorter transcript and gene lengths observed for lncRNAs compared to both coding transcript sets indicates that reduced length is likely an intrinsic property of sponge lncRNAs, rather than an artifact introduced during transcriptome assembly.

Exon and intron lengths were more similar between lncRNAs and PCTs compared to transcript and gene lengths. This indicates that the shorter size of sponge lncRNAs is driven primarily by reduced exon number rather than by uniformly shorter exon or intron units. The lower isoform diversity may reflect simpler lncRNA gene architecture, lower expression levels that reduce isoform detection power, or stronger filtering against ambiguous transcript models. These

explanations are not mutually exclusive. Because alternative isoforms are more difficult to assemble and validate for lowly expressed genes (Grabherr et al., 2011), observed isoform counts may underestimate the true isoform diversity of sponge lncRNAs.

5.7. Sponge Long Non-coding RNAs Show Predominantly Canonical Splicing but Reduced Splice-Junction Support

Splice-site analysis showed that confirmed lncRNA introns were dominated by canonical GT–AG splice sites, similar to introns from assembly-derived and BRAKER3-derived PCGs. Sequence logos of canonical splice junctions were highly similar across all three transcript classes, indicating that confirmed sponge lncRNA introns use broadly conserved splice signals. This finding supports the reliability of the lncRNA catalogue, as it suggests that read-supported lncRNA introns are not dominated by random assembly artefacts but instead display recognizable splice-site architecture.

However, the proportion of lncRNA splice junctions supported by RNA-seq reads was substantially lower than that observed for assembly- or BRAKER3-derived PCGs. Several factors may contribute to this pattern. First, lncRNA genes are typically expressed at lower levels than PCGs (Statello et al., 2021; Mattick et al., 2023), resulting in reduced read coverage across splice junctions and consequently lower confidence in junction validation. Low expression levels may also increase the susceptibility of lncRNA transcript assemblies to fragmentation and inaccuracies in exon boundary prediction. In addition, genome quality may influence splice-junction support, particularly in repetitive or incompletely assembled regions where transcript-to-genome alignments are less reliable. Some sponge species also exhibited elevated proportions of non-canonical GC–AG and AT–AC splice junctions within lncRNA introns. These observations should be interpreted with caution. While these junctions are recognised minor alternatives to canonical GT–AG splice sites and may therefore represent genuine non-canonical splicing events (Turunen et al., 2013), enrichment of these signals could alternatively reflect assembly artifacts or errors in splice-junction inference.

Splice-site strength was further evaluated to distinguish well-supported canonical junctions from weaker or potentially non-canonical splice signals. The lower scores observed for non-canonical junctions were expected, as MaxEntScan models are primarily optimized for canonical splice-site features and therefore exhibit reduced sensitivity to divergent or rare splice signals. In addition, because MaxEntScan was developed using human splice-site data, absolute score values should be interpreted cautiously when applied to evolutionarily distant taxa such as sponges (Yeo & Burge, 2004). Nevertheless, the relative scoring pattern was biologically plausible. Overall, these findings support the presence of conserved canonical splicing in sponge lncRNAs, while emphasizing that weakly supported and non-canonical junctions require cautious interpretation.

5.8. *Sponge Long Non-coding RNAs Differ from Coding Transcripts in Sequence Composition and Exhibit Diverse Genomic Organization*

K-mer-based sequence composition analysis generally separated lncRNAs from both assembly-derived and BRAKER3-derived PCTs. This compositional distinction likely reflects the absence of protein-coding constraints in lncRNA transcripts. Protein-coding sequences are shaped by codon usage, amino acid constraints, reading-frame preservation, and selection against disruptive mutations (Yang & Nielsen, 2008). By contrast, lncRNAs are not constrained by protein-coding potential, although their sequences may still be shaped by RNA secondary structure, regulatory motifs, TE insertions, and local genomic context (Statello et al., 2021; Mattick et al., 2023; Kapusta et al., 2013). The generally lower GC content of lncRNAs in most species is consistent with these compositional differences, although the magnitude of GC-content divergence was modest.

A notable exception was *S. ciliatum*, in which lncRNAs and assembly-derived PCTs exhibited almost complete overlap in compositional space. This pattern likely reflects the highly fragmented nature of the *S. ciliatum* transcriptome assembly, which may have reduced the distinction between coding and non-coding transcript sets by generating large numbers of incomplete transcript fragments.

The distribution of lncRNA genomic classes varied substantially among species, but several broad patterns were consistent. Intergenic lncRNAs represented the dominant class in most species, whereas bidirectional lncRNAs were consistently rare. Antisense lncRNAs were abundant in some species, particularly *A. beatrice* and *E. muelleri*, while intronic lncRNAs were enriched in *H. caerulea* and *S. ciliatum*. These differences may partly reflect biological variation in genome organization and transcriptional regulation, but they may also be influenced by annotation quality and RNA-seq sampling. Because lncRNA classification depends on BRAKER3-derived PCG annotations, incomplete or fragmented coding gene models can alter the assignment of lncRNA classes.

The predominance of intergenic lncRNAs observed in most sponge species is consistent with studies from a wide range of eukaryotes. For instance, large-scale lncRNA annotation studies in several kingdoms have reported that intergenic transcripts constitute the majority of identified lncRNAs, whereas bidirectional and intronic classes are generally less abundant. Similarly, antisense lncRNAs frequently represent the second largest class and can comprise a substantial fraction of the lncRNA repertoire, reflecting the pervasive occurrence of antisense transcription in eukaryotic genomes. These patterns have been observed in vertebrates (GENCODE, 2025; Darbellay & Necseulea, 2020), fungi (Avalos et al., 2025), plants (Jha et al., 2020), and sponges (Gaiti et al., 2015; Bodulić, 2020).

Genomic class was also associated with differences in lncRNA structure. Overlapping lncRNAs generally exhibited the longest gene bodies and introns as their loci intersect extended protein-coding regions or complex transcriptional units. Their increased isoform number in

several species may also reflect more complex local transcription, or alternatively assembly ambiguity in gene-dense regions. Intronic and bidirectional lncRNAs were generally more compact, consistent with their more constrained genomic positions. Intergenic and antisense lncRNAs occupied intermediate structural ranges. These patterns suggest that genomic context may influence lncRNA gene architecture, either through biological regulation or assembly complexity.

5.9. *Transposable Elements Are Associated with Sponge Long Non-coding RNAs*

Across most analysed sponge species, TEs were enriched in lncRNA regions relative to PCGs. This pattern was particularly evident for DNA transposons and LTR retrotransposons, especially in exonic regions. The enrichment of TE-derived sequence in lncRNA genes is consistent with studies in vertebrates showing that TEs contribute to lncRNA origin, diversification, and regulatory potential (Mattick et al., 2023; Hezroni et al., 2015; Kapusta et al., 2013). Similar observations have also been found in sponges (Gaiti et al., 2015; Bodulić, 2020). Notably, TEs can provide promoters, splice sites, polyadenylation signals, RNA-binding motifs, and structural domains. Because lncRNAs are not constrained by ORFs, they may tolerate TE insertions more readily than protein-coding exons (Mattick et al., 2023; Hezroni et al., 2015; Kapusta et al., 2013). This likely explains why differences between lncRNAs and PCGs were particularly strong in exonic regions.

The enrichment analysis relative to genomic background provided a more nuanced interpretation. Although lncRNA genes contained more TE-derived sequence than PCGs, lncRNA exons were often depleted for TE sequence relative to the genome-wide abundance of each repeat class. By contrast, lncRNA introns were more frequently enriched. This pattern suggests that lncRNA exons are not neutrally accumulating TE sequence, but remain subject to constraints related to transcript processing, stability, or function. Intronic regions, in contrast, appear to tolerate repeat accumulation more readily. Species-specific enrichment patterns likely reflect differences in repeat landscape, genome size, assembly quality, and lineage-specific TE activity. For example, elevated LTR enrichment in lncRNA introns of *O. lobularis*, *P. ficiformis*, *S. lacustris*, and *S. ciliatum* suggests that retrotransposons may have contributed differently to lncRNA evolution in these lineages. Furthermore, SINEs and helitrons were less abundant overall but showed strong enrichment in selected species, indicating that low-abundance repeat classes can contribute disproportionately to lncRNA-associated regions.

5.10. *Sponge Long Non-coding RNAs Show Limited Sequence and Syntenic Conservation*

Pairwise sequence comparisons revealed generally limited conservation among sponge lncRNAs, with most species pairs sharing fewer than 10% of lncRNA transcripts. This estimate should be interpreted in the context of the stringent BLASTn parameters applied in this analysis, which were designed to prioritise high-confidence sequence similarities rather than maximise the detection of potential homologues. The low level of sequence conservation is

therefore consistent both with the rapid sequence evolution commonly observed for lncRNAs (Hezroni et al., 2015; Huang et al., 2024; Quinn et al., 2016) and with the deep phylogenetic divergence among the analysed sponge species. The higher levels of sequence similarity observed between more closely related or ecologically comparable species pairs, including *E. muelleri* and *S. lacustris*, *C. crambe* and *T. wilhelma*, and *D. avara* and *S. officinalis*, indicate that phylogenetic distance is a significant determinant of detectable lncRNA conservation.

Sequence-similarity clusters conserved across three or more species were relatively rare. Most clusters contained lncRNAs from three or four species, whereas clusters spanning larger numbers of species were uncommon and were characterized by longer alignments with higher gap frequencies. This pattern suggests that broadly conserved lncRNAs are difficult to align across distant sponge lineages and may retain only short conserved regions rather than full-length transcript similarity. Antisense lncRNAs were strongly enriched among class-conserved sequence clusters, possibly reflecting constraints imposed by overlap with conserved PCGs or by local regulatory relationships. By contrast, intergenic lncRNAs were underrepresented among class-conserved sequence clusters despite representing the most abundant lncRNA class overall, implying a rapid turnover of intergenic lncRNA sequences.

Synteny analysis revealed an additional layer of lncRNA conservation. Because lncRNAs may retain conserved genomic positions despite substantial sequence divergence, synteny-based approaches can identify candidate conserved regulatory loci that are not detectable by sequence-similarity searches (Hezroni et al., 2015; Huang et al., 2024; Quinn et al., 2016). The criteria applied here were intentionally stringent. Syntenic lncRNAs were required to occur within conserved PCG neighbourhoods and to satisfy class-specific contextual criteria. This reduced the likelihood of false-positive syntenic assignments, but likely underestimated the full extent of lncRNA positional conservation, particularly in fragmented genome assemblies.

Syntenic conservation was generally limited, but was strongest between *E. muelleri* and *S. lacustris*, where conserved lncRNAs were distributed across multiple chromosome-scale syntenic blocks. Only a subset of syntenic lncRNAs also showed direct sequence similarity to their syntenic counterparts, indicating that positional conservation and sequence conservation capture overlapping but distinct aspects of lncRNA evolution. Broadly conserved syntenic neighbourhoods were especially informative. The neighbourhood containing *ARMC2*, *FOXJ1*, *CFAP46*, *CBY1*, *LACE1*, and *FANCI* was conserved across five species and included multiple intergenic lncRNAs. Several of these genes are associated with ciliary function (Mukherjee et al., 2019), suggesting that conserved lncRNAs may occur near genes involved in core sponge cellular processes. Similarly, the *TGFB3*-associated neighbourhood contained numerous conserved lncRNAs near genes involved in developmental regulation, chromatin remodelling, ribosome biogenesis, and splicing. These regions therefore represent strong candidates for future functional investigation.

Enrichment analyses of PCGs associated with conserved lncRNAs suggested possible links to phosphorylation, protein folding, cytoskeletal organization, ciliary functions, and signal transduction. These enrichments should be interpreted cautiously, as they are based on neighbouring or overlapping PCGs rather than direct functional characterization of the lncRNAs themselves. In addition, GO annotations are derived from bilaterian model organisms and may not translate directly to sponge biology. Nevertheless, the recurrence of related functional themes across species and conservation categories suggests that conserved sponge lncRNAs may preferentially occur near genes involved in development, cellular organization, and environmental response. The enrichment of dynein- and microtubule-related molecular functions is particularly notable given the importance of ciliated and flagellated cell types in sponge biology (Simpson, 1984; Leys & Hill, 2012).

5.11. *Sponge Long Non-coding RNA Genes Are Lowly Expressed and Sample-Specific*

Expression analyses showed that lncRNA genes were expressed at significantly lower levels than PCGs in every analysed RNA-seq library. This represented one of the most consistent patterns observed in the dissertation and is concordant with findings from sponges and other animal species (Gaiti et al., 2015; Fortunato et al., 2014; Bodulić, 2020; Statello et al., 2021; Mattick et al., 2023). Low lncRNA expression may reflect rapid transcript turnover or regulatory functions that require only limited transcript abundance. Although the expression distributions of lncRNA genes and PCGs partially overlapped, the overall lncRNA gene expression distribution was strongly shifted toward lower TPM values. Notably, this difference is likely conservative, as the lncRNA identification pipeline excluded transcripts with TPM < 1, introducing a bias toward the retention of more highly expressed lncRNA candidates.

Furthermore, lncRNA expression was highly sample-dependent. Many lncRNAs were detected in only a small fraction of libraries, whereas comparatively few were expressed across most or all samples. This pattern is in accordance with highly context-specific lncRNA expression across biological conditions, tissues, cell types, and developmental stages (Statello et al., 2021; Mattick et al., 2023). Finally, no consistent expression differences were observed among lncRNA genomic classes, although intergenic and bidirectional lncRNAs occasionally showed broader expression ranges. As such, genomic class alone was not a reliable predictor of lncRNA expression level.

5.12. *Long Non-coding RNAs as Putative Regulators of Sponge Immune Responses*

The analyses presented thus far primarily focused on the identification and annotation of sponge lncRNAs. To investigate their potential functional relevance, a publicly available dataset examining MAMP exposure in *A. aerophoba* and *D. avara* was reanalysed as a case study of transcriptional responses to microbial immune stimulation (Pita et al., 2018). Differential expression analysis showed that both lncRNA genes and PCGs responded to MAMP exposure. In *A. aerophoba*, the transcriptional response increased over time, with the largest number of DEGs detected at 5 h post-exposure and a predominance of upregulation.

In *D. avara*, the response was more balanced, with lncRNA genes showing a stronger tendency toward repression at 5 h. These species-specific observations may reflect differences in immune strategy, microbiome composition, or physiological response to the same stimulus. Importantly, many lncRNAs showed large fold changes, indicating that sponge lncRNAs are not merely static transcriptional by-products, but can respond strongly to environmental and microbial cues.

Additionally, WGCNA provided a network-level view of these responses by identifying co-expression modules associated with MAMP treatment. The selected modules contained both lncRNA genes and PCGs, and their eigengene profiles captured distinct temporal patterns, including progressive induction, early induction, and sustained repression. Enrichment analysis of PCGs within these modules linked MAMP-responsive networks to signalling, immune-related processes, cell-cycle regulation, translation, and extracellular matrix regulation. These enrichments are consistent with a coordinated response to microbial stimulation and suggest that lncRNAs are embedded within functionally relevant transcriptional programs.

Hub lncRNAs identified within MAMP-responsive modules represent the strongest candidate regulatory transcripts from this analysis. Several hub lncRNAs were differentially expressed and exhibited high module membership scores, indicating that their expression closely followed module-level dynamics. Particularly notable were intronic hub lncRNAs in greenyellow modules of both species, which were strongly upregulated after MAMP exposure and showed high module membership. These transcripts may represent candidate regulators or markers of MAMP-responsive gene networks. Other highly connected hub lncRNAs were not differentially expressed, indicating that network centrality and differential expression capture distinct aspects of potential biological relevance. Although this case study does not demonstrate direct regulatory function, it provides a prioritised set of candidate lncRNAs for future experimental investigation. Functional validation could include targeted knockdown, spatial expression profiling, co-localisation with neighbouring or co-expressed PCGs, and time-course experiments using additional immune stimuli. More broadly, these results support the hypothesis that sponge lncRNAs participate in dynamic gene regulatory responses and may contribute to host–microbe interactions in early-branching animals.

6. CONCLUSION

This dissertation provides an integrated methodological and comparative framework for the identification and analysis of lncRNAs in sponges. By combining transcriptome assembly assessment, genome decontamination, genome annotation, and comparative lncRNA analyses, the study addressed key challenges associated with lncRNA discovery in non-model holobiont organisms. The main conclusions are:

- The CATS framework enables reliable transcriptome assembly evaluation in non-model organisms by combining reference-free and reference-based approaches.
- CATS-rf provides interpretable transcript-level quality scores based on RNA-seq read support and outperforms existing reference-free tools across most simulated and realistic benchmarks.
- CATS-rb complements CATS-rf by estimating transcriptome completeness and structural consistency through transcript-to-genome alignment without requiring existing genome annotation.
- Joint interpretation of CATS-rf and CATS-rb offers full insight into transcriptome assembly quality by capturing both transcript-level accuracy and transcriptome completeness.
- The combined use of de novo, genome-guided, and clustering-based assembly strategies represents an effective approach for maximizing transcriptome quality and completeness in non-model organisms such as sponges.
- No single assembly strategy performed optimally across all sponge species, demonstrating the importance of systematic quality assessment before downstream lncRNA discovery.
- Genome decontamination is an important prerequisite for sponge genomic analyses considering that sponge assemblies frequently contain contaminant sequences derived from associated microorganisms.
- The decontamination pipeline developed in this dissertation effectively removed contaminant scaffolds while preserving host genome completeness, providing a cleaner genomic foundation for lncRNA characterization.
- Repeat and PCG annotation revealed substantial genomic heterogeneity among sponge species, including major differences in TE composition and predicted gene number.
- The lncRNA identification workflow recovered extensive candidate lncRNA repertoires across all analysed sponge species, substantially expanding the known lncRNA catalogue in Porifera.
- Differences in lncRNA repertoire size among species are strongly influenced by RNA-seq sampling, library availability, sequencing depth, and biological context.
- Sponge lncRNAs share major structural characteristics with lncRNAs from other animals, including shorter transcript and gene lengths, fewer exons, and lower isoform diversity compared with PCTs.

- Sponge lncRNAs are predominantly canonically spliced, although they show reduced splice-junction support relative to PCTs, likely reflecting greater assembly uncertainty.
- Sponge lncRNAs differ from PCTs in sequence composition and GC content, supporting their distinction as a separate transcript class.
- Intergenic lncRNAs represent the dominant genomic class in most sponge species.
- Transposable elements are strongly associated with sponge lncRNAs, particularly in intronic regions, suggesting that TEs have contributed to lncRNA origin and diversification.
- Sponge lncRNAs show limited sequence conservation across species, consistent with rapid lncRNA evolution and deep phylogenetic divergence within Porifera.
- Synteny-based analyses identified a smaller set of positionally conserved lncRNAs, including candidates located near genes involved in ciliary function, development, chromatin regulation, and cellular organization.
- Sponge lncRNAs are generally lowly expressed and highly sample-specific, consistent with patterns observed in other animals and supporting their potential roles in context-dependent regulation.
- MAMP-responsive lncRNAs in *A. aerophoba* and *D. avara* demonstrate that sponge lncRNAs can respond dynamically to microbial immune stimulation.
- Co-expression network analyses identified candidate hub lncRNAs embedded in immune- and signalling-associated transcriptional modules, suggesting potential regulatory roles in host–microbe interactions.
- The methodological framework developed here provides a foundation for future comparative studies of lncRNAs in early-branching animals and other poorly annotated non-model organisms.

7. REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi:10.1016/S0022-2836(05)80360-2.
- Antoszewski, K., Chmielewska, K., Jagiello, K. and Puzyn, T. (2025) 'Designing RNA sequencing experiments: a practical guide to reproducible gene expression analysis', *Computational and Structural Biotechnology Journal*, 31, pp. 101–119. doi:10.1016/j.csbj.2025.12.015.
- Arrigoni, A., Ranzani, V., Rossetti, G., Panzeri, I., Abrignani, S., Bonnal, R.J. and Pagani, M. (2016) 'Analysis of RNA-seq and noncoding RNA', *Methods in Molecular Biology*, 1480, pp. 125–135. doi:10.1007/978-1-4939-6380-5_11.
- Astashyn, A., Tvedte, E.S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strobe, P.K., Sylla, P.M., Wagner, L., Bidwell, S.L., Brown, L.C., Clark, K., Davis, E.W., Smith-White, B., Hlavina, W., Pruitt, K.D., Schneider, V.A. and Murphy, T.D. (2024) 'Rapid and sensitive detection of genome contamination at scale with FCS-GX', *Genome Biology*, 25(1), 60. doi:10.1186/s13059-024-03198-7.
- Atianand, M.K. and Fitzgerald, K.A. (2014) 'Long non-coding RNAs and control of gene expression in the immune system', *Trends in Molecular Medicine*, 20(11), pp. 623–631. doi:10.1016/j.molmed.2014.06.002.
- Aubry, S., Kelly, S., Kümpers, B.M., Smith-Unna, R.D. and Hibberd, J.M. (2014) 'Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis', *PLoS Genetics*, 10(6), e1004365. doi:10.1371/journal.pgen.1004365.
- Avalos, J., Perera-Bonaño, A. and Limón, M.C. (2025) 'Identification and functions of lncRNAs in fungi', *Non-Coding RNA*, 11(5), 72. doi:10.3390/ncrna11050072.
- Bálint, B., Merényi, Z., Hegedüs, B., Grigoriev, I.V., Hou, Z., Földi, C. and Nagy, L.G. (2024) 'ContScout: sensitive detection and removal of contamination from annotated genomes', *Nature Communications*, 15(1), 936. doi:10.1038/s41467-024-45024-5.
- Bhan, A., Soleimani, M. and Mandal, S.S. (2017) 'Long noncoding RNA and cancer: a new paradigm', *Cancer Research*, 77(15), pp. 3965–3981. doi:10.1158/0008-5472.CAN-16-2634.
- Bodulić, K. and Vlahoviček, K. (2026) 'Comprehensive assessment of transcriptome assembly quality using CATS', *Nature Communications*. doi:10.1038/s41467-026-72171-8.
- Bodulić, K. and Vlahoviček, K. (2026) 'Hidden Contaminants in Sponge Genomes: Large-Scale Decontamination of 30 Public Assemblies', *bioRxiv*. doi:10.1101/2025.07.22.666112

- Bodulić, K. (2020) Computational analysis of long non-coding RNAs of the cave sponge *Eunapius subterraneus*. Master's thesis. University of Zagreb, Faculty of Science, Department of Biology. Available at: <https://urn.nsk.hr/urn:nbn:hr:217:310016> (Accessed: 12 April 2026).
- Botting, J.P. and Muir, L.A. (2018) 'Early sponge evolution: a review and phylogenetic framework', *Palaeoworld*, 27(1), pp. 1–29. doi:10.1016/j.palwor.2017.07.001.
- Botting, J.P., Zhang, Y. and Muir, L.A. (2017) 'Discovery of missing link between demosponges and hexactinellids confirms palaeontological model of sponge evolution', *Scientific Reports*, 7(1), 5286. doi:10.1038/s41598-017-05604-6.
- Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*, 34(5), pp. 525–527. doi:10.1038/nbt.3519.
- Briggs, J.A., Wolvetang, E.J., Mattick, J.S., Rinn, J.L. and Barry, G. (2015) 'Mechanisms of long non-coding RNAs in mammalian nervous system development, plasticity, disease, and evolution', *Neuron*, 88(5), pp. 861–877. doi:10.1016/j.neuron.2015.09.045.
- Buchfink, B., Xie, C. and Huson, D.H. (2015) 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods*, 12(1), pp. 59–60. doi:10.1038/nmeth.3176.
- Bushmanova, E., Antipov, D., Lapidus, A. and Prjibelski, A.D. (2019) 'rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data', *GigaScience*, 8(9), giz100. doi:10.1093/gigascience/giz100.
- Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V. and Prjibelski, A.D. (2016) 'rnaQUAST: a quality assessment tool for de novo transcriptome assemblies', *Bioinformatics*, 32(14), pp. 2210–2212. doi:10.1093/bioinformatics/btw218.
- Cabau, C., Escudié, F., Djari, A., Guiguen, Y., Bobe, J. and Klopp, C. (2017) 'Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies', *PeerJ*, 5, e2988. doi:10.7717/peerj.2988.
- Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021) 'eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale', *Molecular Biology and Evolution*, 38(12), pp. 5825–5829. doi:10.1093/molbev/msab293.
- Chen, N. (2004) 'Using RepeatMasker to identify repetitive elements in genomic sequences', *Current Protocols in Bioinformatics*, Chapter 4, Unit 4.10. doi:10.1002/0471250953.bi0410s05.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. and Mortazavi, A. (2016) 'A survey of best

practices for RNA-seq data analysis', *Genome Biology*, 17, 13. doi:10.1186/s13059-016-0881-8.

Cotto, K.C., Feng, Y.Y., Ramu, A., Richters, M., Freshour, S.L., Skidmore, Z.L., Xia, H., McMichael, J.F., Kunisaki, J., Campbell, K.M., Chen, T.H., Rozycki, E.B., Adkins, D., Devarakonda, S., Sankararaman, S., Lin, Y., Chapman, W.C., Maher, C.A., Arora, V., Dunn, G.P., Uppaluri, R., Govindan, R., Griffith, O.L. and Griffith, M. (2023) 'Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer', *Nature Communications*, 14(1), 1589. doi:10.1038/s41467-023-37266-6.

Darbellay, F. and Necsulea, A. (2020) 'Comparative transcriptomics analyses across species, organs, and developmental stages reveal functionally constrained lncRNAs', *Molecular Biology and Evolution*, 37(1), pp. 240–259. doi:10.1093/molbev/msz212.

Dawadi, P., Pokharel, B., Shrestha, A., Niraula, D., Naeem, A., Miura, S., Roy, M. and Nepal, S. (2025) 'From bench to bytes: a practical guide to RNA sequencing data analysis', *Frontiers in Genetics*, 16, 1697922. doi:10.3389/fgene.2025.1697922.

Degnan, B.M., Vervoort, M., Larroux, C. and Richards, G.S. (2009) 'Early evolution of metazoan transcription factors', *Current Opinion in Genetics & Development*, 19(6), pp. 591–599. doi:10.1016/j.gde.2009.09.008.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J. and Guigó, R. (2012) 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Research*, 22(9), pp. 1775–1789. doi:10.1101/gr.132159.111.

Diedenhofen, B. and Musch, J. (2015) 'cocor: a comprehensive solution for the statistical comparison of correlations', *PLoS One*, 10(3), e0121945. doi:10.1371/journal.pone.0121945.

Dunn, C.W., Hejnal, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sørensen, M.V., Haddock, S.H.D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q. and Giribet, G. (2008) 'Broad phylogenomic sampling improves resolution of the animal tree of life', *Nature*, 452(7188), pp. 745–749. doi:10.1038/nature06614.

Eddy, S.R. (2011) 'Accelerated profile HMM searches', *PLoS Computational Biology*, 7(10), e1002195. doi:10.1371/journal.pcbi.1002195.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F.A. (2020) 'RepeatModeler2 for automated genomic discovery of transposable element families',

Proceedings of the National Academy of Sciences of the United States of America, 117(17), pp. 9451–9457. doi:10.1073/pnas.1921046117.

Fortunato, S.A.V., Adamski, M., Bergum, B., Guder, C., Jordal, S., Leininger, S., Zwafink, C., Rapp, H.T. and Adamska, M. (2014) 'Regulatory RNA at the root of animals: dynamic expression of developmental lincRNAs in the calcsponge *Sycon ciliatum*', *Genome Biology and Evolution*, 6(6), pp. 1400–1414. doi:10.1093/gbe/evu104.

Francis, W.R., Eitel, M., Vargas, S., Adamski, M., Haddock, S.H.D., Krebs, S., Blum, H. and Wörheide, G. (2017) 'The genome of the contractile demosponge *Tethya wilhelma* and the evolution of metazoan neural signalling pathways', *bioRxiv*. doi:10.1101/120998.

Francis, W.R., Eitel, M., Vargas, S., Garcia-Escudero, C.A., Conci, N., Deister, F., Mah, J.L., Guiglielmoni, N., Krebs, S., Blum, H., Leys, S.P. and Wörheide, G. (2023) 'The genome of the reef-building glass sponge *Aphrocallistes vastus* provides insights into silica biomineralization', *Royal Society Open Science*, 10(6), 230423. doi:10.1098/rsos.230423.

Frazee, A.C., Jaffe, A.E., Langmead, B. and Leek, J.T. (2015) 'Polyester: simulating RNA-seq datasets with differential transcript expression', *Bioinformatics*, 31(17), pp. 2778–2784. doi:10.1093/bioinformatics/btv272.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152. doi:10.1093/bioinformatics/bts565.

Gabriel, L., Brūna, T., Hoff, K.J., Ebel, M., Lomsadze, A., Borodovsky, M. and Stanke, M. (2024) 'BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA', *Genome Research*, 34(5), pp. 769–777. doi:10.1101/gr.278090.123.

Gaiti, F., Fernandez-Valverde, S.L., Nakanishi, N., Calcino, A.D., Yanai, I., Tanurdzic, M. and Degnan, B.M. (2015) 'Dynamic and widespread lncRNA expression in a sponge and the origin of animal complexity', *Molecular Biology and Evolution*, 32(9), pp. 2367–2382. doi:10.1093/molbev/msv117.

Gazave, E., Lapébie, P., Renard, E., Vacelet, J., Rocher, C., Ereskovsky, A.V., Lavrov, D.V. and Borchellini, C. (2010) 'Molecular phylogeny restores the supra-generic subdivision of homoscleromorph sponges (Porifera, Homoscleromorpha)', *PLoS One*, 5(12), e14290. doi:10.1371/journal.pone.0014290.

Gotoh, O. (2024) 'Spaln3: improvement in speed and accuracy of genome mapping and spliced alignment of protein query sequences', *Bioinformatics*, 40(8), btae517. doi:10.1093/bioinformatics/btae517.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652. doi:10.1038/nbt.1883.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P. and Ulitsky, I. (2015) 'Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species', *Cell Reports*, 11(7), pp. 1110–1122. doi:10.1016/j.celrep.2015.04.023.

Hentschel, U., Piel, J., Degnan, S.M. and Taylor, M.W. (2012) 'Genomic insights into the marine sponge microbiome', *Nature Reviews Microbiology*, 10(9), pp. 641–654. doi:10.1038/nrmicro2839.

Hölzer, M. and Marz, M. (2019) 'De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers', *GigaScience*, 8(5), giz039. doi:10.1093/gigascience/giz039.

Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., Edger, P.P., Altman, N.S., Pires, J.C., Leebens-Mack, J.H. and dePamphilis, C.W. (2016) 'Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome', *PLoS ONE*, 11(1), e0146062. doi:10.1371/journal.pone.0146062.

Huang, W., Xiong, T., Zhao, Y., Heng, J., Han, G., Wang, P., Zhao, Z., Shi, M., Li, J., Wang, J., Wu, Y., Liu, F., Xi, J.J., Wang, Y. and Zhang, Q.C. (2024) 'Computational prediction and experimental validation identify functionally conserved lncRNAs from zebrafish to human', *Nature Genetics*, 56, pp. 124–135. doi:10.1038/s41588-023-01620-7.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A. and Wheeler, T.J. (2016) 'The Dfam database of repetitive DNA families', *Nucleic Acids Research*, 44(D1), pp. D81–D89. doi:10.1093/nar/gkv1272.

Huson, D.H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D. and Williams, R.B.H. (2018) 'MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs', *Biology Direct*, 13(1), 6. doi:10.1186/s13062-018-0208-7.

Jackson, D.J., Cerveau, N. and Posnien, N. (2024) 'De novo assembly of transcriptomes and differential gene expression analysis using short-read data from emerging model organisms: a brief guide', *Frontiers in Zoology*, 21(1), 17. doi:10.1186/s12983-024-00538-y.

Jha, U.C., Nayyar, H., Jha, R., Khurshid, M., Zhou, M., Mantri, N. and Siddique, K.H.M. (2020) 'Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation', *BMC Plant Biology*, 20(1), 466. doi:10.1186/s12870-020-02595-x.

- Jing, D., Mei, Y., Zhou, H., Huang, C., Wang, Z., Jin, S., Chen, H., Duanmu, H., Liu, Y., Li, F. and He, K. (2025) 'The landscape and evolutionary dynamics of derived lncRNAs in insects', *Communications Biology*, 9(1), 85. doi:10.1038/s42003-025-09347-3.
- Kang, Y.J., Yang, D.C., Kong, L., Hou, M., Meng, Y.Q., Wei, L. and Gao, G. (2017) 'CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features', *Nucleic Acids Research*, 45(W1), pp. W12–W16. doi:10.1093/nar/gkx428.
- Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M. and Feschotte, C. (2013) 'Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs', *PLoS Genetics*, 9(4), e1003470. doi:10.1371/journal.pgen.1003470.
- Katoh, K. and Standley, D.M. (2013) 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi:10.1093/molbev/mst010.
- Kazazian, H.H. Jr. (2004) 'Mobile elements: drivers of genome evolution', *Science*, 303(5664), pp. 1626–1632. doi:10.1126/science.1089670.
- Kelly, J.B., Carlson, D.E., Low, J.S. and Thacker, R.W. (2022) 'Novel trends of genome evolution in highly complex tropical sponge microbiomes', *Microbiome*, 10(1), 164. doi:10.1186/s40168-022-01359-z.
- Kenny, N.J., Francis, W.R., Rivera-Vicéns, R.E., Juravel, K., de Mendoza, A. and Díez-Vives, C. (2020) 'Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*', *Nature Communications*, 11, 3676. doi:10.1038/s41467-020-17397-w.
- Kent, W.J. (2002) 'BLAT—the BLAST-like alignment tool', *Genome Research*, 12(4), pp. 656–664. doi:10.1101/gr.229202.
- Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nature Biotechnology*, 37(8), pp. 907–915. doi:10.1038/s41587-019-0201-4.
- Kung, J.T., Colognori, D. and Lee, J.T. (2013) 'Long noncoding RNAs: past, present, and future', *Genetics*, 193(3), pp. 651–669. doi:10.1534/genetics.112.146704.
- Kuzman, M. (2020) Identification and characterization of transposable elements in sponges (Porifera). Doctoral thesis. University of Zagreb, Faculty of Science, Department of Biology. Available at: <https://urn.nsk.hr/urn:nbn:hr:217:082440> (Accessed: 5 June 2026).

- Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., Grant, G.R. and Hogenesch, J.B. (2014) 'IVT-seq reveals extreme bias in RNA sequencing', *Genome Biology*, 15(6), R86. doi:10.1186/gb-2014-15-6-r86.
- Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, 9, 559. doi:10.1186/1471-2105-9-559.
- Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi:10.1038/nmeth.1923.
- Larroux, C., Fahey, B., Liubicich, D., Hinman, V.F., Gauthier, M., Gongora, M., Green, K., Wörheide, G., Leys, S.P. and Degnan, B.M. (2006) 'Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity', *Evolution & Development*, 8(2), pp. 150–173. doi:10.1111/j.1525-142X.2006.00086.x.
- Lataretu, M., Krautwurst, S., Huska, M.R., Marquet, M., Viehweger, A., Braun, S.D., Brandt, C. and Hölzer, M. (2025) 'Targeted decontamination of sequencing data with CLEAN', *NAR Genomics and Bioinformatics*, 7(3), lqaf105. doi:10.1093/nargab/lqaf105.
- Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) 'Software for computing and annotating genomic ranges', *PLoS Computational Biology*, 9(8), e1003118. doi:10.1371/journal.pcbi.1003118.
- Leininger, S., Adamski, M., Bergum, B., Guder, C., Liu, J., Lapébie, P., Rapp, H.T., Adamska, M. and Degnan, B.M. (2014) 'Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans', *Nature Communications*, 5, 3905. doi:10.1038/ncomms4905.
- Leinonen, R., Sugawara, H., Shumway, M. and International Nucleotide Sequence Database Collaboration (2011) 'The Sequence Read Archive', *Nucleic Acids Research*, 39(Database issue), pp. D19–D21. doi:10.1093/nar/gkq1019.
- Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. and Regev, A. (2010) 'Comprehensive comparative analysis of strand-specific RNA sequencing methods', *Nature Methods*, 7(9), pp. 709–715. doi:10.1038/nmeth.1491.
- Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R. and Pfister, H. (2014) 'UpSet: visualization of intersecting sets', *IEEE Transactions on Visualization and Computer Graphics*, 20(12), pp. 1983–1992. doi:10.1109/TVCG.2014.2346248.
- Leys, S.P. and Hill, A. (2012) 'The physiology and molecular biology of sponge tissues', *Advances in Marine Biology*, 62, pp. 1–56.

- Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R. and Dewey, C.N. (2014) 'Evaluation of de novo transcriptome assemblies from RNA-Seq data', *Genome Biology*, 15(12), 553. doi:10.1186/s13059-014-0553-5.
- Li, W. and Jiang, T. (2012) 'Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads', *Bioinformatics*, 28(22), pp. 2914–2921. doi:10.1093/bioinformatics/bts559.
- Li, X., Kong, Y., Zhao, Q.-Y., Li, G., Zhang, Y., Wang, J. and Wang, J. (2016) 'De novo assembly of transcriptome from next-generation sequencing data', *Quantitative Biology*, 4, pp. 94–105. doi:10.1007/s40484-016-0069-y.
- Liao, Y., Smyth, G.K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30(7), pp. 923–930. doi:10.1093/bioinformatics/btt656.
- Liu, Y., Ding, W., Yu, W., Zhang, Y., Ao, X. and Wang, J. (2021) 'Long non-coding RNAs: biogenesis, functions, and clinical significance in gastric cancer', *Molecular Therapy: Oncolytics*, 23, pp. 458–476. doi:10.1016/j.omto.2021.11.005.
- Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), 550. doi:10.1186/s13059-014-0550-8.
- Maldonado, M., Pita, L., Erpenbeck, D., Hentschel, U., Oatley, G., Sinclair, E., Aunin, E., Gettle, N., Santos, C., Paulini, M., Niu, H., McKenna, V. and O'Brien, R. (2025) 'The chromosomal genome sequence of the sponge *Corticium candelabrum* Schmidt, 1862 and its associated microbial metagenome sequences', *Wellcome Open Research*, 10, 535. doi:10.12688/wellcomeopenres.24923.1.
- Manuel, M., Borchellini, C., Alivon, E., Le Parco, Y., Vacelet, J. and Boury-Esnault, N. (2003) 'Phylogeny and evolution of calcareous sponges: monophyly of *Calcinea* and *Calcaronea*, high level of morphological homoplasy, and the primitive nature of axial symmetry', *Systematic Biology*, 52(3), pp. 311–333. doi:10.1080/10635150390196966.
- Martin, J.A. and Wang, Z. (2011) 'Next-generation transcriptome assembly', *Nature Reviews Genetics*, 12(10), pp. 671–682. doi:10.1038/nrg3068.
- Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10–12. doi:10.14806/ej.17.1.200.
- Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M.E., Fitzgerald, K.A., Gingeras, T.R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J.B., Lee, J.T., Mendell, J.T., Mercer, T.R., Moore, K.J.,

Nakagawa, S., Rinn, J.L., Spector, D.L., Ulitsky, I., Wan, Y., Wilusz, J.E. and Wu, M. (2023) 'Long non-coding RNAs: definitions, functions, challenges and recommendations', *Nature Reviews Molecular Cell Biology*, 24(6), pp. 430–447. doi:10.1038/s41580-022-00566-8.

Metzker, M.L. (2010) 'Sequencing technologies — the next generation', *Nature Reviews Genetics*, 11(1), pp. 31–46. doi:10.1038/nrg2626.

Milligan, M.J. and Lipovich, L. (2015) 'Pseudogene-derived lncRNAs: emerging regulators of gene expression', *Frontiers in Genetics*, 5, 476. doi:10.3389/fgene.2014.00476.

Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., Ptitsyn, A., Reshetov, D., Mukherjee, K., Moroz, T.P., Bobkova, Y., Yu, F., Kapitonov, V.V., Jurka, J., Bobkov, Y.V., Swore, J.J., Girardo, D.O., Fodor, A., Gusev, F., Sanford, R., Bruders, R., Kittler, E., Mills, C.E., Rast, J.P., Derelle, R., Solovyev, V.V., Kondrashov, F.A., Swalla, B.J., Sweedler, J.V., Rogaev, E.I., Halanych, K.M. and Kohn, A.B. (2014) 'The ctenophore genome and the evolutionary origins of neural systems', *Nature*, 510(7503), pp. 109–114. doi:10.1038/nature13400.

Morrow, C. and Cárdenas, P. (2015) 'Proposal for a revised classification of the Demospongiae (Porifera)', *Frontiers in Zoology*, 12, 7. doi:10.1186/s12983-015-0099-8.

Nam, J.W. and Bartel, D.P. (2012) 'Long noncoding RNAs in *C. elegans*', *Genome Research*, 22(12), pp. 2529–2540. doi:10.1101/gr.140475.112.

Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) 'Infernal 1.0: inference of RNA alignments', *Bioinformatics*, 25(10), pp. 1335–1337. doi:10.1093/bioinformatics/btp157.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., Jiang, N., Hirsch, C.N. and Hufford, M.B. (2019) 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome Biology*, 20(1), 275. doi:10.1186/s13059-019-1905-y.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H. and Soldatov, A. (2009) 'Transcriptome analysis by strand-specific sequencing of complementary DNA', *Nucleic Acids Research*, 37(18), e123. doi:10.1093/nar/gkp596.

Patterson, J., Carpenter, E.J., Zhu, Z., An, D., Liang, X., Geng, C., Drmanac, R. and Wong, G.K. (2019) 'Impact of sequencing depth and technology on de novo RNA-Seq assembly', *BMC Genomics*, 20(1), 604. doi:10.1186/s12864-019-5965-x.

Pawlik, J.R. and McMurray, S.E. (2020) 'The emerging ecological and biogeochemical importance of sponges on coral reefs', *Annual Review of Marine Science*, 12, pp. 315–337. doi:10.1146/annurev-marine-010419-010807.

Peng, P.H., Hsu, K.W., Lai, J.C.-Y. and Wu, K.J. (2021) 'The role of hypoxia-induced long noncoding RNAs (lncRNAs) in tumorigenesis and metastasis', *Biomedical Journal*, 44(5), pp. 521–533. doi:10.1016/j.bj.2021.03.005.

Peng, Y., Leung, H.C.M., Yiu, S.M., Lv, M.J., Zhu, X.G. and Chin, F.Y.L. (2013) 'IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels', *Bioinformatics*, 29(13), pp. i326–i334. doi:10.1093/bioinformatics/btt219.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. and Salzberg, S.L. (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology*, 33(3), pp. 290–295. doi:10.1038/nbt.3122.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D.J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. and Manuel, M. (2009) 'Phylogenomics revives traditional views on deep animal relationships', *Current Biology*, 19(8), pp. 706–712. doi:10.1016/j.cub.2009.02.052.

Pita, L., Hoepfner, M.P., Ribes, M. and Hentschel, U. (2018) 'Differential expression of immune receptors in two marine sponges upon exposure to microbial-associated molecular patterns', *Scientific Reports*, 8(1), 16081. doi:10.1038/s41598-018-34330-w.

Pita, L., Rix, L., Slaby, B.M., Franke, A. and Hentschel, U. (2018) 'The sponge holobiont in a changing ocean: from microbes to ecosystems', *Microbiome*, 6(1), 46. doi:10.1186/s40168-018-0428-1.

Quinn, J.J., Zhang, Q.C., Georgiev, P., Ilik, I.A., Akhtar, A. and Chang, H.Y. (2016) 'Rapid evolutionary turnover underlies conserved lncRNA–genome interactions', *Genes & Development*, 30(2), pp. 191–207. doi:10.1101/gad.272187.115.

Raghavan, V., Kraft, L., Mesny, F. and Rigerte, L. (2022) 'A simple guide to de novo transcriptome assembly and annotation', *Briefings in Bioinformatics*, 23(2), bbab563. doi:10.1093/bib/bbab563.

Ramakrishnaiah, Y., Kuhlmann, L. and Tyagi, S. (2020) 'Towards a comprehensive pipeline to identify and functionally annotate long noncoding RNA (lncRNA)', *Computers in Biology and Medicine*, 127, 104028. doi:10.1016/j.combiomed.2020.104028.

Redmond, A.K. and McLysaght, A. (2021) 'Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding', *Nature Communications*, 12, 1783. doi:10.1038/s41467-021-22074-7.

Renard, E., Leys, S.P., Wörheide, G. and Borchiellini, C. (2018) 'Understanding animal evolution: the added value of sponge transcriptomics and genomics: the disconnect between

gene content and body plan evolution', *BioEssays*, 40(9), e1700237. doi:10.1002/bies.201700237.

Robbins, S.J., Song, W., Engelberts, J.P., Glasl, B., Slaby, B.M., Boyd, J., Marangon, E., Botté, E.S., Laffy, P., Thomas, T. and Webster, N.S. (2021) 'A genomic view of the microbiome of coral reef demosponges', *The ISME Journal*, 15, pp. 1641–1654. doi:10.1038/s41396-020-00850-6.

Salehi, S., Taheri, M.N., Azarpira, N., Zare, A. and Behzad-Behbahani, A. (2017) 'State of the art technologies to explore long non-coding RNAs in cancer', *Journal of Cellular and Molecular Medicine*, 21(12), pp. 3120–3140. doi:10.1111/jcmm.13238.

Sayers, E.W., Bolton, E.E., Fine, A.M., Kelly, C., Kim, S., Landrum, M., Lathrop, S., Malheiro, A., Murphy, T.D., Phan, L., Pujar, S., Trawick, B.W., Schneider, V.A. and Pruitt, K.D. (2026) 'Database resources of the National Center for Biotechnology Information in 2026', *Nucleic Acids Research*, 54(D1), pp. D20–D27. doi:10.1093/nar/gkaf1060.

Scrucca, L., Fraley, C., Murphy, T.B. and Raftery, A.E. (2023) Model-based clustering, classification, and density estimation using mclust in R. Boca Raton: Chapman and Hall/CRC. doi:10.1201/9781003277965.

Seppey, M., Manni, M. and Zdobnov, E.M. (2019) 'BUSCO: assessing genome assembly and annotation completeness', *Methods in Molecular Biology*, 1962, pp. 227–245. doi:10.1007/978-1-4939-9173-0_14.

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y. and Ge, Q. (2021) 'Bias in RNA-seq library preparation: current challenges and solutions', *Biomed Research International*, 2021, 6647597. doi:10.1155/2021/6647597.

Simpson, T.L. (1984) The cell biology of sponges. New York: Springer.

Singhal, S. (2013) 'De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set', *Molecular Ecology Resources*, 13, pp. 403–416. doi:10.1111/1755-0998.12077.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M. and Kelly, S. (2016) 'TransRate: reference-free quality assessment of de novo transcriptome assemblies', *Genome Research*, 26(8), pp. 1134–1144. doi:10.1101/gr.196469.115.

Song, L. and Florea, L. (2015) 'Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads', *GigaScience*, 4, 48. doi:10.1186/s13742-015-0089-y.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N.H., Stanke, M., Adamska, M., Darling, A., Degnan, S.M., Oakley, T.H., Plachetzki, D.C., Zhai, Y., Adamski, M., Calcino, A.,

Cummins, S.F., Goodstein, D.M., Harris, C., Jackson, D.J., Leys, S.P., Shu, S., Woodcroft, B.J., Vervoort, M., Kosik, K.S., Manning, G., Degnan, B.M. and Rokhsar, D.S. (2010) 'The Amphimedon queenslandica genome and the evolution of animal complexity', *Nature*, 466(7307), pp. 720–726. doi:10.1038/nature09201.

Statello, L., Guo, C.-J., Chen, L.-L. and Huarte, M. (2021) 'Gene regulation by long non-coding RNAs and its biological functions', *Nature Reviews Molecular Cell Biology*, 22(2), pp. 96–118. doi:10.1038/s41580-020-00315-9.

Steffen, K., Proux-Wéra, E., Soler, L., Churcher, A., Sundh, J. and Cárdenas, P. (2023) 'Whole genome sequence of the deep-sea sponge *Geodia barretti* (Metazoa, Porifera, Demospongiae)', *G3: Genes, Genomes, Genetics*, 13(10), jkad192. doi:10.1093/g3journal/jkad192.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) 'Differential expression in RNA-seq: a matter of depth', *Genome Research*, 21(12), pp. 2213–2223. doi:10.1101/gr.124321.111.

Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E.M. and Kriventseva, E.V. (2025) 'OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes', *Nucleic Acids Research*, 53(D1), pp. D516–D522. doi:10.1093/nar/gkae987.

Thomas, J. and Pritham, E.J. (2015) 'Helitrons, the eukaryotic rolling-circle transposable elements', *Microbiology Spectrum*, 3(4), MDNA3-0049-2014. doi:10.1128/microbiolspec.MDNA3-0049-2014.

Thomas, T., Moitinho-Silva, L., Lurgi, M., Björk, J.R., Easson, C., Astudillo-García, C., Olson, J.B., Erwin, P.M., López-Legentil, S., Luter, H., Chaves-Fonnegra, A., Costa, R., Schupp, P.J., Steindler, L., Erpenbeck, D., Gilbert, J., Knight, R., Ackermann, G., Lopez, J.V., Taylor, M.W., Thacker, R.W., Montoya, J.M., Hentschel, U. and Webster, N.S. (2016) 'Diversity, structure and convergent evolution of the global sponge microbiome', *Nature Communications*, 7, 11870. doi:10.1038/ncomms11870.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols*, 7(3), pp. 562–578. doi:10.1038/nprot.2012.016.

Turunen, J.J., Niemelä, E.H., Verma, B. and Frilander, M.J. (2013) 'The significant other: splicing by the minor spliceosome', *WIREs RNA*, 4(1), pp. 61–76. doi:10.1002/wrna.1141.

Ungaro, A., Pech, N., Martin, J.F., McCairns, R.J.S., Mévy, J.P., Chappaz, R. and Gilles, A. (2017) 'Challenges and advances for transcriptome assembly in non-model species', *PLoS ONE*, 12(9), e0185020. doi:10.1371/journal.pone.0185020.

- Wang, S. and Gribskov, M. (2017) 'Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis', *Bioinformatics*, 33(3), pp. 327–333. doi:10.1093/bioinformatics/btw625.
- Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., Kissinger, J.C. and Paterson, A.H. (2012) 'MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity', *Nucleic Acids Research*, 40(7), e49. doi:10.1093/nar/gkr1293.
- Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature Reviews Genetics*, 10(1), pp. 57–63. doi:10.1038/nrg2484.
- Wei, C.W., Luo, T., Zou, S.S. and Wu, A.S. (2018) 'The role of long noncoding RNAs in central nervous system and neurodegenerative diseases', *Frontiers in Behavioral Neuroscience*, 12, 175. doi:10.3389/fnbeh.2018.00175.
- Wickham, H. (2016) *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
- Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20(1), 257. doi:10.1186/s13059-019-1891-0.
- Wörheide, G., Dohrmann, M., Erpenbeck, D., Larroux, C., Maldonado, M., Voigt, O., Borchellini, C. and Lavrov, D.V. (2012) 'Deep phylogeny and evolution of sponges (phylum Porifera)', *Advances in Marine Biology*, 61, pp. 1–78. doi:10.1016/B978-0-12-387778-1.00007-6.
- Wörheide, G., Francis, W.R., Deister, F., Krebs, S., Erpenbeck, D. and Vargas, S. (2024) 'The genomes of the aquarium sponges *Tethya wilhelma* and *Tethya minuta* (Porifera: Demospongiae)', *F1000Research*, 13, 679. doi:10.12688/f1000research.150836.2.
- Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X. and Yu, G. (2021) 'clusterProfiler 4.0: a universal enrichment tool for interpreting omics data', *The Innovation*, 2(3), 100141. doi:10.1016/j.xinn.2021.100141.
- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.W., Li, Y., Xu, X., Wong, G.K. and Wang, J. (2014) 'SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads', *Bioinformatics*, 30(12), pp. 1660–1666. doi:10.1093/bioinformatics/btu077.
- Yan, H., Bombarely, A. and Li, S. (2020) 'DeepTE: a computational method for de novo classification of transposons with convolutional neural network', *Bioinformatics*, 36(15), pp. 4269–4275. doi:10.1093/bioinformatics/btaa519.

Yang, Z. and Nielsen, R. (2008) 'Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage', *Molecular Biology and Evolution*, 25(3), pp. 568–579. doi:10.1093/molbev/msm284.

Yeo, G. and Burge, C.B. (2004) 'Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals', *Journal of Computational Biology*, 11(2–3), pp. 377–394. doi:10.1089/1066527041410418.

Zhang, R.G., Li, G.Y., Wang, X.L., Dainat, J., Wang, Z.X., Ou, S. and Ma, Y. (2022) 'TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes', *Horticulture Research*, 9, uhac017. doi:10.1093/hr/uhac017.

Web sites

Alexa, A. and Rahnenfuhrer, J. (2025) topGO: Enrichment analysis for gene ontology. Available at: <https://bioconductor.org/packages/topGO> (Accessed: 3 June 2026).

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Schwendinger, B. and Krylov, I. (2025) data.table: Extension of data.frame. Available at: <https://CRAN.R-project.org/package=data.table> (Accessed: 30 April 2026).

Bembom, O. and Ivanek, R. (2026) seqLogo: Sequence logos for DNA sequence alignments. Available at: <https://bioconductor.org/packages/seqLogo> (Accessed: 26 May 2026).

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D. and Müller, K. (2026) igraph: Network analysis and visualization in R. Available at: <https://CRAN.R-project.org/package=igraph> (Accessed: 3 June 2026).

Gao, C. and Dusa, A. (2025) ggVennDiagram: A ggplot2 implementation of Venn diagrams. Available at: <https://CRAN.R-project.org/package=ggVennDiagram> (Accessed: 2 May 2026).

GENCODE (2025) GENCODE Release 49. Available at: <https://www.genencodegenes.org> (Accessed: 17 April 2026).

Kassambara, A. (2023) ggcorrplot: Visualization of a correlation matrix using ggplot2. Available at: <https://CRAN.R-project.org/package=ggcorrplot> (Accessed: 30 April 2026).

Kolde, R. (2025) pheatmap: Pretty heatmaps. R package version 1.0.13. Available at: <https://CRAN.R-project.org/package=pheatmap> (Accessed: 30 April 2026).

Konopka, T. (2023) umap: Uniform manifold approximation and projection. Available at: <https://CRAN.R-project.org/package=umap> (Accessed: 26 May 2026).

Li, B. (2014) RSEM-EVAL GitHub repository. Available at: <https://github.com/nfillmore/rsem-eval> (Accessed: 23 April 2026).

Madeira, P. (2016) TransRate GitHub repository branch. Available at: <https://github.com/pmomadeira/transrate> (Accessed: 30 April 2026).

Miles, A. (2013) pysamstats: Read-level statistics for BAM files. Available at: <https://github.com/alimanfoo/pysamstats> (Accessed: 28 April 2025).

New England Biolabs (2022) NEBNext® Ultra™ RNA Library Prep Kit for Illumina (E7530). Available at: <https://www.neb.com/en/products/e7530-nebnext-ultra-rna-library-prep-kit-for-illumina> (Accessed: 14 April 2026).

Pagès, H., Aboyoun, P., Gentleman, R. and DebRoy, S. (2025) Biostrings: Efficient manipulation of biological strings. Available at: <https://bioconductor.org/packages/Biostrings> (Accessed: 30 April 2026).

Pedersen, T. (2025) ggraph: An implementation of grammar of graphics for graphs and networks. Available at: <https://CRAN.R-project.org/package=ggraph> (Accessed: 4 June 2026).

R Core Team (2026) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/> (Accessed: 30 April 2026).

Sherrill-Mix, S. (2025) taxonomizr: Functions to work with NCBI accessions and taxonomy. Available at: <https://CRAN.R-project.org/package=taxonomizr> (Accessed: 30 April 2026).

Signorell, A. (2025) DescTools: Tools for descriptive statistics. Available at: <https://CRAN.R-project.org/package=DescTools> (Accessed: 30 April 2026).

Smith-Unna, R. (2016) TransRate GitHub repository. Available at: <https://github.com/blahah/transrate> (Accessed: 23 April 2026).

Stephens, M., Carbonetto, P., Gerard, D., Lu, M., Sun, L., Willwerscheid, J. and Xiao, N. (2023) ashR: Methods for adaptive shrinkage using empirical Bayes. Available at: <https://CRAN.R-project.org/package=ashr> (Accessed: 3 June 2026).

van den Boogaart, K.G., Tolosana-Delgado, R. and Bren, M. (2025) compositions: Compositional data analysis. Available at: <https://CRAN.R-project.org/package=compositions> (Accessed: 30 April 2026).

Vu, V.Q. and Friendly, M. (2024) ggbiplot: A grammar of graphics implementation of biplots. Available at: <https://CRAN.R-project.org/package=ggbiplot> (Accessed: 30 April 2026).

Wilke, C. (2025) cowplot: Streamlined plot themes and plot annotations for ggplot2. Available at: <https://CRAN.R-project.org/package=cowplot> (Accessed: 30 April 2026).

8. APPENDIX

8.1. *CATS Code and Dataset Availability*

The CATS framework is implemented in R and Bash. The source code is publicly available under the MIT license in the following GitHub repositories:

CATS-rf: <https://github.com/bodulic/CATS-rf>

CATS-rb: <https://github.com/bodulic/CATS-rb>

Both repositories are distributed under the MIT license.

The code used for benchmarking and evaluation of the CATS framework is available at:

https://github.com/bodulic/CATS_benchmark

Benchmark results used for figure generation are provided as standalone tables and are publicly accessible through Zenodo:

<https://doi.org/10.5281/zenodo.16837970>

8.2. *Genome Decontamination Pipeline Code and Dataset Availability*

The genome decontamination pipeline is publicly available under the MIT license on GitHub:

https://github.com/bodulic/genome_decontamination

The code used for pipeline benchmarking and genome assembly decontamination is available on Github:

https://github.com/bodulic/sponge_genome_decontamination_workflow

The decontaminated versions of all sponge genome assemblies generated in this study are publicly available through Zenodo:

<https://doi.org/10.5281/zenodo.19064896>

Each assembly is provided in FASTA format together with MD5 checksum files. For every analysed genome assembly, accompanying reports include per-scaffold contamination statistics, contaminant taxonomic assignments, and BUSCO completeness assessments. Results from the validation analyses are also included in the repository.

8.3. *Long Non-coding RNA Analysis*

The code used for RNA-seq quality control, transcriptome assembly, lncRNA identification, and comparative analyses is publicly available on GitHub:

https://github.com/bodulic/sponge_lncrna

The repository also contains details on the utilized RNA-seq libraries, along with quality metrics of all assembled transcriptomes.

The lncRNA datasets generated in this study are publicly available through Zenodo:

<https://doi.org/10.5281/zenodo.20634426>.

The repository contains the complete set of identified lncRNA transcripts for all analysed sponge species. In addition, detailed genomic annotation files (GFF3 format) are provided for each lncRNA, including genomic coordinates, exon boundaries, and classification information.

9. CURRICULUM VITAE

Kristian Bodulić is a bioinformatics scientist and PhD candidate in Biology at the Faculty of Science, University of Zagreb, specializing in transcriptomics, next-generation sequencing data analysis, and statistical modelling. Since November 2024, he has served as Head of the Bioinformatics and Statistics Department at the University Hospital for Infectious Diseases “Dr. Fran Mihaljević” in Zagreb, where he leads computational biology activities supporting research and clinical projects. Prior to this role, he worked as a Bioinformatics Scientist at the same institution from 2020 to 2024. Since 2022, he has also held the position of Adjunct Assistant at the Biology Division of the Faculty of Science, University of Zagreb, where he participates in graduate-level teaching and conducts independent and collaborative research.

He obtained a BSc in Molecular Biology (*summa cum laude*) in 2018 and an MSc in Molecular Biology (*magna cum laude*) in 2020 from the Faculty of Science, University of Zagreb. His master's thesis focused on the computational analysis of long non-coding RNAs in the endemic cave sponge *Eunapius subterraneus*. He is the lead developer of the CATS (Comprehensive Assessment of Transcript Sequences) framework, comprising the CATS-rf and CATS-rb tools for reference-free and reference-based transcriptome assembly quality assessment. This work was published in *Nature Communications* in 2026 and represents a significant contribution to transcriptome analysis in non-model organisms. He is also the lead developer of an integrated genome decontamination pipeline designed to identify and remove contaminating sequences from holobiont genomes, facilitating reliable genomic analyses in non-model species.

Throughout his career, he has co-authored 28 peer-reviewed scientific publications in the fields of bioinformatics, transcriptomics, virology, infectious diseases, and molecular epidemiology, including five publications as first author. He has also served as a reviewer for several international scientific journals. He received the Rector's Award of the University of Zagreb in 2019 and the Dean's Award for Excellence from the Faculty of Science in 2018. During his studies, he participated in several science outreach and public engagement activities organized by the Faculty of Science. He is a native Croatian speaker and is proficient in English (C1) and Italian (B2).