Genomic characterization of long-noncoding RNAs in the zebrafish genome

Genomiczna charakterystyka długich niekodujących RNA w genomie danio pręgowanego

Monika Kwiatkowska

PhD thesis executed in

the Department of Computational Biology of Non-coding RNA
at the Institute of Bioorganic Chemistry Polish Academy of Sciences

Poznań Doctoral School of the Institutes of the Polish Academy of Sciences

Thesis Supervisor:
Dr. hab. Barbara Uszczyńska-Ratajczak, prof. IBCH PAS

Co-supervisor:
Dr. Sílvia Carbonell Sala

Poznań 2024

This work and the scholarship of Monika Kwiatkowska were funded by:

National Science Centre Opus grant 2018/31/B/NZ2/01940
Principal Investigator: Dr. hab. Barbara Uszczyńska-Ratajczak

National Science Centre Sonata Bis grant 2021/42/E/NZ2/00434
Principal Investigator: Dr. hab. Barbara Uszczyńska-Ratajczak

PUBLICATIONS

Barbara Uszczynska-Ratajczak, Sreedevi Sugunan, **Monika Kwiatkowska**, Maciej Migdal, Silvia Carbonell-Sala, Anna Sokol, Cecilia L Winata, Agnieszka Chacinska. 2022. “Profiling subcellular localization of nuclear-encoded mitochondrial gene products in zebrafish”. *Life Sci Alliance*. 2022;6(1):e202201514.

doi:10.26508/lsa.202201514

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Dr. hab. Barbara Uszczyńska-Ratajczak, and my co-supervisor, Dr. Sílvia Carbonell Sala, for granting me the opportunity to contribute to such an extraordinary project. Their unwavering support, guidance, and understanding—both on professional and personal levels—have been pivotal in the successful completion of this thesis. I am profoundly grateful for their insightful scientific discussions, constructive feedback, and the thoughtful advice they have provided at every stage of my research. Their mentorship has served as a great source of inspiration for me in various facets of scientific inquiry.

I would like to express my sincere gratitude to all current and former members of the Department of Computational Biology of Non-coding RNA for their assistance and engaging discussions. In particular, I am grateful to Tomasz Mądry, Marta Blangiewicz, Sasti Das and Daniel Kuźnicki for fruitful discussions and assistance in bioinformatic analysis whenever I needed it. I would like to extend special thanks to Agata Chmielewska and Katarzyna Solka for their invaluable administrative support. I would not have been able to accomplish anything without your assistance.

I would like to extend my special thanks to the current and former members of the Laboratory of Zebrafish Developmental Genomics, particularly Agata Sulej, Maciej Łapiński, and Eugeniusz Tralle. Their invaluable suggestions and guidance in teaching me new experimental techniques significantly contributed to the progress of my work.

I am appreciative of Professor Jacek Kuźnicki for providing me with the opportunity to volunteer in his lab, which was instrumental in enabling me to complete the experimental work related to my PhD project. I am immensely grateful to Tomasz Węgierski for sharing his expertise in microscopy imaging, which greatly enriched my knowledge and experience.

I would like to express my sincere gratitude to all the members of the Zebrafish Core Facility for their invaluable support and assistance. Their provision of zebrafish biological samples was crucial to the successful completion of my PhD project.

I am deeply grateful for the knowledge and experience I gained during my internships with Professor Roderic Guigó's group at the Centre for Genomic Regulation. I would like to extend my special thanks to Carme Arnan Ros, Silvia Pérez Lluch, and Marina Ruiz Romero for their invaluable assistance in preparing my sequencing libraries and for the enjoyable time spent together in Barcelona. Additionally, I would like to thank Emilio Palumbo and Gazaldeep Kaur for their support in analyzing the sequencing data.

I am grateful to Professor Rory Johnson and Carlos Pulido for their unwavering support during the selection of targets for functional characterization. Their insightful comments and suggestions provided valuable guidance, helping me navigate this critical aspect of my research.

No one has been more important to me in the completion of this thesis than my husband, Damian. I am deeply grateful for his patience, invaluable assistance, and enduring encouragement. His belief in me provided strength during the most challenging times, inspiring me to persevere. This achievement is as much his as it is mine, as his understanding and support have been vital to the successful realization of this work.

My heartfelt thanks also go to our little girl, Oliwia, for her patience, and resilience at such a young age. Her presence, radiant smile, and unconditional love have been a constant source of strength throughout this process.

I would also like to extend special thanks to my mother-in-law for her invaluable help in caring for our daughter during my absence. Her support allowed me to focus on completing this thesis, and for that, I am deeply grateful.

TABLE OF CONTENTS

ABBREVIATIONS

CAGE - Cap Analysis of Gene Expression

cDNA - complementary DNA

CLS - Capture Long-read Sequencing

CNE - Constructive Neutral Evolution

ConnectOR - Connect Orthologous RNAs

CPAT - Coding Potential Assessment Tool

CPC - Coding Potential Calculator

DNA - deoxyribonucleic acid

dpf - days post fertilization

ENCODE - The Encyclopedia of DNA Elements

ERCC- External RNA Controls Consortium

FANTOM - Functional Annotation of the Mammalian genome

FL - Full-length

FLC - FLOWERING LOCUS C

FOR - Forward

gDNA - genomic DNA

HAVANA - Human and Vertebrate Analysis and Annotation

HCGM - High Confidence Genome Mapping

hpf - hours post fertilization

LA PCR - Long and Accurate Polymerase Chain Reaction

lincRNA - long intergenic non-coding RNA

lncRNA - long non-coding RNA

MANE - Matched Annotation from NCBI and EMBL-EBI

miRNA - microRNA

MIR-HG - miRNA Host Gene

misc_RNA - miscellaneous RNA

mRNA - messenger ribonucleic acid

ncRNA - non-coding RNA

ONT - Oxford Nanopore Technologies

ORF - Open Reading Frame

piRNA - piwi-interacting RNA

PBS - Phosphate-Buffered Saline

PCR - Polymerase Chain Reaction

PFA - Paraformaldehyde

PLAR - Pipeline for LncRNA Annotation from RNA-seq data

RBP - RNA Binding Protein

REV - Reverse

RIN - RNA Integrity Number

RNA - ribonucleic acid

rRNA - ribosomal RNA

RT - Room Temperature

SE - Super Enhancer

siRNA - small interfering RNA

SIRV- Spike-In RNA Variant

SMART - Switching Mechanism At RNA Termini

SN - Supernatant

snRNA - small nuclear RNA

snoRNA - small nucleolar RNA

SS500 - Size-selection 500 bp

SSC - Saline-sodium citrate

TAE - Tris-Acetate-EDTA

TE - Transposable Elements

TE - Typical Enhancer

TGS - Third-Generation Sequencing

TM - Transcript Model

TSS - Transcription Start Site

TTS - Transcription Termination Site

tRNA - transfer RNA

TSO - Template Switching Oligo

UCNE - Ultraconserved Noncoding Elements

UMI - Unique Molecular Identifier

VEGA - Vertebrate Genome Annotation

XCI - X Chromosome Inactivation

ABSTRACT

Vertebrate genomes produce thousands of long noncoding RNAs (lncRNAs) - transcripts longer than 200 nucleotides with limited protein-coding potential. Despite a growing number of lncRNAs involved in crucial biological processes, over 97% of them remain functionally uncharacterized. Employment of animal models can aid in understanding the biological roles of lncRNAs, but the success of this exploration heavily depends on the quality of genome annotations. While zebrafish (*Danio rerio*) has emerged as a powerful and promising vertebrate model for exploring lncRNA biology, its genome annotation lags far behind that of humans or mice, significantly hindering its application. As part of this project, I aimed to create a comprehensive and accurate catalog of lncRNA genes in the zebrafish genome. To achieve this, I optimized the CapTrap-CLS protocol in zebrafish. CapTrap-seq is a full-length library preparation method that has demonstrated superiority in specifically enriching for 5' and 3'-complete transcripts while simultaneously and efficiently reducing the presence of ribosomal RNA (rRNA) molecules. The implementation of CapTrap-seq in zebrafish led to a more accurate annotation of transcription start sites (TSSs) for biologically relevant genes. Furthermore, to enhance the detection of lowly represented lncRNAs, I combined CapTrap-seq with the Capture Long-read Sequencing (CLS) approach—a targeted RNA sequencing method that integrates RNA capture with long-read Oxford Nanopore Technologies (ONT) sequencing. The RNA capture probes effectively enriched lncRNAs, significantly enhancing their representation in the post-capture libraries and extending the reference annotation for the zebrafish genome. Moreover, targeted RNA sequencing of human-mouse-zebrafish syntenic regions not only revealed new transcript isoforms for potentially functional lncRNAs, but also significantly enhanced the detection of novel genes in the intergenic space. This result underscores the importance of positional conservation as an effective strategy for the discovery of novel, potentially functional lncRNA loci. The improved zebrafish genome annotation developed during my PhD project offers a strong foundation for advancing the biological relevance of zebrafish as a model organism for studying the function of lncRNAs.

STRESZCZENIE

Genomy kręgowców kodują tysiące długich niekodujących RNA (lncRNA) — transkryptów o długości przekraczającej 200 nukleotydów, które mają ograniczony potencjał kodowania białek. Pomimo rosnącej liczby lncRNA zaangażowanych w kluczowe procesy biologiczne, ponad 97% z nich wciąż nie ma przypisanej funkcji. Modele zwierzęce mogą odegrać istotną rolę w poznaniu biologicznych funkcji lncRNA, jednak powodzenie tych badań w dużej mierze zależy od jakości adnotacji genomu. Chociaż danio pręgowany (*Danio rerio*) stał się obiecującym modelem kręgowców do badania biologii lncRNA, jakość adnotacji jego genomu wciąż pozostaje znacznie gorsza w porównaniu z adnotacjami genów u ludzi i myszy, co istotnie ogranicza jego zastosowanie. Zadaniem tego projektu było opracowanie kompleksowego i precyzyjnego katalogu genów lncRNA w genomie danio pręgowanego. W tym celu zoptymalizowano protokół CapTrap-CLS. Metoda CapTrap-seq, która służy do przygotowywania bibliotek pełnej długości, wykazuje przewagę w specyficznym wzbogacaniu o transkrypty z kompletnymi końcami 5' i 3', jednocześnie skutecznie redukując ilość rybosomalnego RNA (rRNA). Zastosowanie CapTrap-seq u danio pręgowanego znacząco poprawiło adnotację miejsc rozpoczęcia transkrypcji (TSS) dla biologicznie istotnych genów. Ponadto, aby zwiększyć detekcję słabo reprezentowanych lncRNA, CapTrap-seq połączono z metodą Capture Long-read Sequencing (CLS) — ukierunkowaną techniką sekwencjonowania RNA, która łączy wychwytywanie RNA z sekwencjonowaniem długich odczytów za pomocą platformy ONT (Oxford Nanopore Technologies). To podejście skutecznie wzbogaciło lncRNA, znacznie zwiększając ich reprezentację w bibliotekach sekwencyjnych, co w efekcie doprowadziło do rozszerzenia referencyjnej adnotacji lncRNA dla genomu danio pręgowanego. Dodatkowo, ukierunkowane sekwencjonowanie RNA z regionów pozycyjnie zachowanych pomiędzy człowiekiem, myszą a danio pręgowanym ujawniło nowe izoformy transkryptów dla potencjalnie funkcjonalnych lncRNA oraz istotnie poprawiło detekcję nowych genów w przestrzeniach międzygenowych. Wynik ten podkreśla znaczenie pozycyjnej zachowawczości jako efektywnej strategii w odkrywaniu nowych, potencjalnie funkcjonalnych cząsteczek lncRNA. Zoptymalizowana adnotacja genomu danio pręgowanego, stworzona w ramach tego projektu doktorskiego, stanowi solidną podstawę do wzmocnienia biologicznej wartości tego gatunku jako modelu zwierzęcego do badań nad funkcją lncRNA.

The research context

The development and survival of all living organisms depends on their ability to store and read genetic information contained within their DNA. These instructions, embedded in the sequence of nucleotide bases, are organized into genes—the fundamental units of the genome—that must be transcribed into various functional forms for cellular use. The link between DNA and RNA was first established in the late 1950s by Elliot Volkin and Lawrence Astrachan, who identified RNA as a DNA-like molecule synthesized from DNA (Volkin and Astrachan, 1956). Additionally, Brachet and Caspersson noted a correlation between RNA levels and the rate of protein synthesis (Brachet, 1942; Caspersson, 1947). These pivotal findings led to the formulation of the central dogma of molecular biology by Francis Crick, who articulated the concept that "DNA makes RNA, and RNA makes protein" (Figure 1.1) (Crick, 1970). In this view, RNA molecules were seen as simple intermediaries between the genetic information stored in DNA and protein production (Cobb, 2015; Cobb, 2017), which led to the belief that the genome was mainly composed of protein-coding genes. Early estimates suggested that the human genome contained around 100,000 protein-coding genes (Liang et al., 2000), driven by the assumption that gene count would align with and reflect the complexity of an organism. At the same time, the non-coding regions of the genome were largely dismissed as nonfunctional, often labeled as "junk" DNA (Ohno, 1972).

The completion of the Human Genome Project reduced the estimate of protein-coding genes in human DNA to approximately 20,000, revealing that coding sequences represent only about 1% of our genomic DNA (Ezkurdia, 2014). This shift in understanding prompted a focus on identifying and annotating the functional elements of the genome. As a result, initiatives like the Encyclopedia of DNA Elements (ENCODE) (Djebali, 2012; ENCODE Project Consortium, 2012) and FANTOM (Carninci, 2005) were launched to create detailed maps of transcription, transcriptional regulator binding sites, chromatin states, and histone modifications across various cell types. These efforts revealed that up to 80% of the human genome can be actively transcribed (Djebali, 2012), highlighting the role of non-protein-coding regions in RNA production. However, despite having the human genome sequence for two decades, the functional potential of these non-protein-coding regions remains largely unexplored, earning them the label of "dark matter" in DNA.

Long noncoding RNAs (lncRNAs)—RNA molecules longer than 200 nucleotides with minimal protein-coding potential—are the largest and most intriguing yet enigmatic class of noncoding RNAs. Estimates suggest that the human genome contains over 100,000 lncRNA genes (Uszczyńska-Ratajczak et al., 2018; Zhao et al., 2021), supporting the validity of the initial estimates of total gene numbers. However, it is lncRNAs, not protein-coding genes, that currently represent the largest class of RNAs. Moreover, the correlation between organismal complexity and the number of lncRNA genes is, to date, stronger than that previously proposed for protein-coding genes (Mattick, 2011; Jandura and Krause, 2017; Lee et al., 2019; Moore and Uchida, 2020).

Despite the identification of hundreds of thousands of lncRNAs in vertebrate genomes, only a small fraction (~2-3%) of these genes have been functionally characterized (Bao et al., 2019; Ma et al., 2019; Li et al., 2023). Functional characterization heavily relies on accurate annotation, and lncRNAs; however, unlike protein-coding genes, remain poorly annotated (Amaral et al., 2023). Our understanding of the sequence-structure relationship for lncRNAs is limited contrary to open reading frames in mRNAs. Additionally, lncRNAs evolve much more rapidly than protein-coding genes (Kutter et al., 2012; Hezroni et al., 2015), making it difficult to use homology or functional information to detect novel lncRNA molecules. Consequently, their annotation largely depends on the detection of transcriptomic evidence. While many lncRNAs have demonstrated biological significance, we still lack the tools to conclusively identify which lncRNAs are truly functional and to fully explore their broad biological potential. Most of the current information about lncRNA functionality is derived from loss-of-function studies on human cell lines or mouse models (Andergassen and Rinn, 2022; Lyu et al., 2023). To fully characterize lncRNA functions, it is essential to study them in other genomes, particularly beyond mammals.

Although zebrafish (*Danio rerio*) has become an important vertebrate model for studying lncRNA evolution and function, its genome annotation is considerably less developed compared to humans or mice, both in terms of the number of identified gene loci and the quality of existing lncRNA models. Furthermore, zebrafish gene catalogs are predominantly biased toward embryonic samples (Ulitsky et al., 2011; Pauli et al., 2012; Dhiman, et al., 2015; Zhou et al., 2024), in contrast to human gene catalogs, which focus on adult tissues and cell lines (Iyer et al., 2015; Szcześniak et al., 2020;

Frankish et al., 2023). As a result, zebrafish lncRNA annotations are incomplete and not fully comprehensive, limiting their comparability to human lncRNA annotation and reducing their overall effectiveness in predicting biological functions of lncRNAs. Therefore, to effectively validate lncRNA functions, it is essential to enhance the use of the zebrafish model. However, this will only be possible with prior improvements in its genome annotation. Enhancing zebrafish genome annotation relies heavily on developing new methods for full-length RNA sequencing and addressing the current bias toward developmental samples. A key priority should be improving the detection of potentially functional lncRNAs, particularly those with positional conservation across genomes, to support their functional analysis. Systematic identification and characterization of evolutionarily conserved lncRNAs is expected to greatly advance our understanding of these functional molecules and how their roles are encoded in their genomic sequences.

1. INTRODUCTION

1.1. Noncoding RNA: Expanding the Central Dogma of Molecular Biology

For many years, RNAs were thought to serve solely as intermediaries or components in protein synthesis, as outlined by the central dogma of molecular biology (Crick, 1958; Cobb, 2015; Cobb, 2017). However, the completion of the Human Genome Project revealed that these processes account for ~1-2% of the genome, leaving the function of the remaining ~99% of noncoding DNA largely unexplored (Lander et al., 2001; Venter et al., 2001). Subsequent discoveries demonstrated that these noncoding regions are actively transcribed, generating a vast array of non-coding RNA species of various types (Djebali, 2012).

Due to the high heterogeneity of these ncRNA transcripts, molecule size is widely used for their classification (Hombach and Kretz, 2016). Consequently, non-coding RNAs are divided into two main classes: small ncRNAs—RNA molecules shorter than 200 nucleotides, and long non-coding RNAs (**lncRNAs**), with a length over 200 base pairs (Figure 1.2). Among small ncRNAs, we can distinguish small nuclear RNAs (**snRNAs**), involved in RNA splicing; small nucleolar RNAs (**snoRNAs**), which are involved in chemical modifications of other RNA molecules; microRNAs (**miRNAs**), Piwi-interacting RNAs (**piRNAs**), and small interfering RNAs (**siRNAs**) that play roles in post-transcriptional regulation by inhibiting translation or promoting the degradation of specific mRNAs (Hombach and Kretz, 2016). Long ncRNAs that are found to be the most numerous group, can be further segregated into linear and circular RNAs (Hombach and Kretz, 2016; Amin et al., 2019). This group has been found to be involved in multiple biological processes, spanning from housekeeping functions such as gene expression regulation to more specialized functions such as genomic imprinting or dosage compensation (Zhang et al., 2019; Statello et al., 2021; Mattick et al., 2023). Altogether, this suggests that non-coding RNAs play an important role in regulating all steps of the central dogma of molecular biology.

It is important to emphasize that the findings of the post-genomic era do not undermine the significance of protein-coding genes or the principles of the central dogma of molecular biology. The central dogma remains a foundational concept for explaining the flow and expression of genetic information in living organisms. However, the discovery of non-coding RNAs (ncRNAs), including long non-coding RNAs (lncRNAs), adds a new layer of complexity to this process, deepening our understanding of gene regulation and cellular function at multiple levels (Figure 1.1).

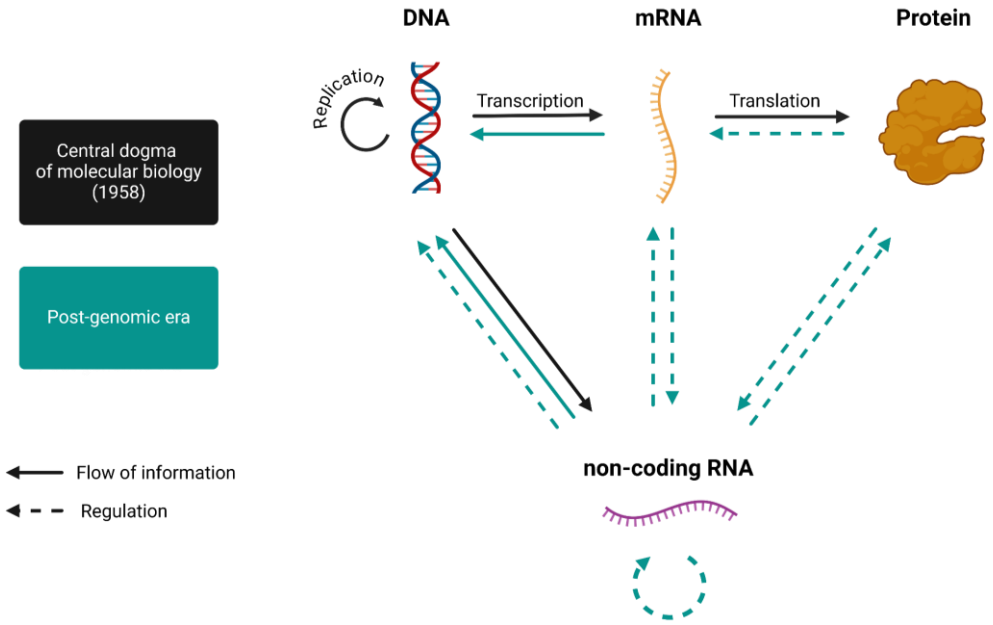


Figure 1.1. The classical dogma of molecular biology and its expanded interpretation in the post-genomic era. *Initial dogma (1958) is represented in black while post-genomic view is shown in dynasty green. Full arrows indicate the flow of genetic information and dashed arrows represent regulatory interactions. Figure based on (Jarroux et al., 2017).*

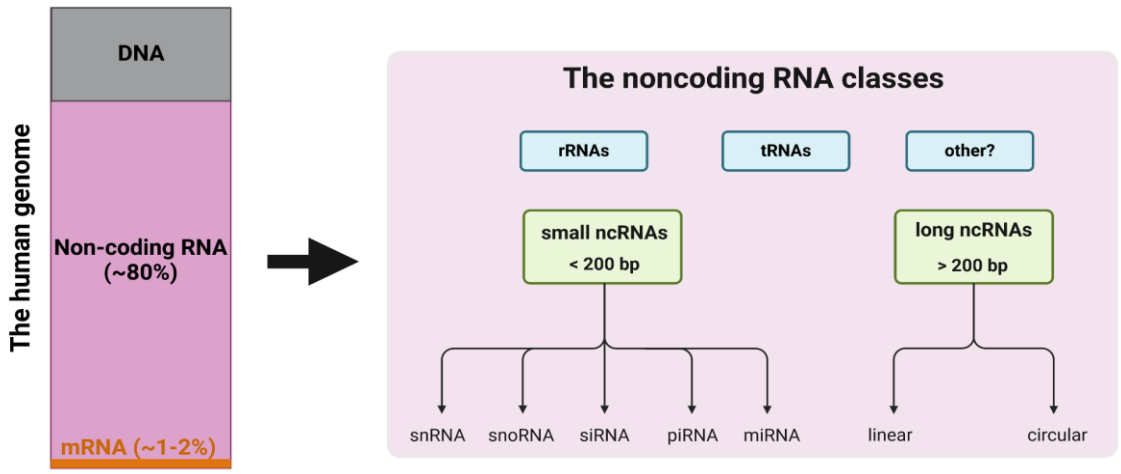


Figure 1.2. The classification of non-protein coding RNA types.

1.2. LncRNAs - the largest class of noncoding RNA molecules

The long non-coding RNAs (lncRNAs) are RNA molecules longer than 200 nucleotides with limited or no protein-coding potential. They represent the most numerous and highly heterogeneous group of genes in the human genome that differ in their genomic origin, biogenesis and mode of action (Statello et al., 2021; Mattick et al., 2023). LncRNAs have been detected across nearly all kingdoms of life, from simple viruses (Wang et al., 2017), bacteria (Harris and Breaker, 2018), and plants (Yadav et al., 2023) to highly complex species (Ulitsky and Bartel, 2013). LncRNA loci form the largest class of genes within the human genome (Kaur et al., 2024). According to the latest GENCODE annotation (v. 47), the human genome encodes 35,934 lncRNA loci, far exceeding the number of protein-coding genes (19,433) and small noncoding RNAs (7,565) (Kaur et al., 2024). A similar trend is observed in other mammalian genomes, including mouse (36,172 lncRNA loci versus 21,470 protein-coding genes, GENCODE annotation vM36) (Kaur et al., 2024). Worldwide annotation efforts are further expanding these estimates to 100,000 and beyond (Uszczyńska-Ratajczak et al., 2018; Ma et al., 2019; Li et al., 2023).

LncRNAs share several similarities with mRNAs: they are typically capped, polyadenylated, and spliced (Mattick et al., 2023; Cabili, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). However, despite these common traits, they perform different roles within the cell, resulting in significant differences in their cellular processing pathways (Statello et al., 2021). Protein-coding genes follow a straightforward expression path: DNA is transcribed into RNA in the nucleus, which is then immediately exported to the cytoplasm for protein synthesis. In contrast, the final product of lncRNA gene expression is a RNA molecule that, due to its unstable nature, must quickly localize to its specific site of action. Consequently, lncRNAs show distinct subcellular localization in the cell. They can either remain nuclear-retained or be directed to various subcellular compartments, including the cytoplasm, mitochondria, or endoplasmic reticulum (ER) (Figure 1.3) (Statello et al., 2021; Bridges et al., 2021). The subcellular localization of long noncoding RNAs (lncRNAs) is closely linked to their functionality, as they are specifically targeted to the sites where they exert their molecular actions (Carlevaro-Fita and Johnson, 2019; Statello et al., 2021; Bridges et al., 2021). Unlike mRNAs, many lncRNAs are primarily localized in the nucleus (Guo et al., 2020; Statello et al., 2021).

Studies have shown that export to the cytoplasm is positively correlated with the splicing efficiency of transcript molecules (Statello et al., 2021, Khan et al., 2023). LncRNAs are generally spliced less efficiently than mRNAs, which explains their increased retention in the nucleus (Figure 1.3) (Khan et al., 2023; Statello et al., 2021).

The primary goal of a protein-coding gene is to produce a functional protein product, leading to high evolutionary conservation of these regions and limiting the number of alternative isoforms (Figure 1.3) (Deveson et al., 2018). In contrast, lncRNAs are not constrained by codon usage, giving them greater flexibility in adapting to mutations or genomic rearrangements (Mattick et al., 2023). They also exhibit highly complex, almost universal splicing patterns. Their exons combinations can be more freely rearranged, resulting in a nearly limitless variety of noncoding isoforms (Deveson et al., 2018; Basu et al., 2023).

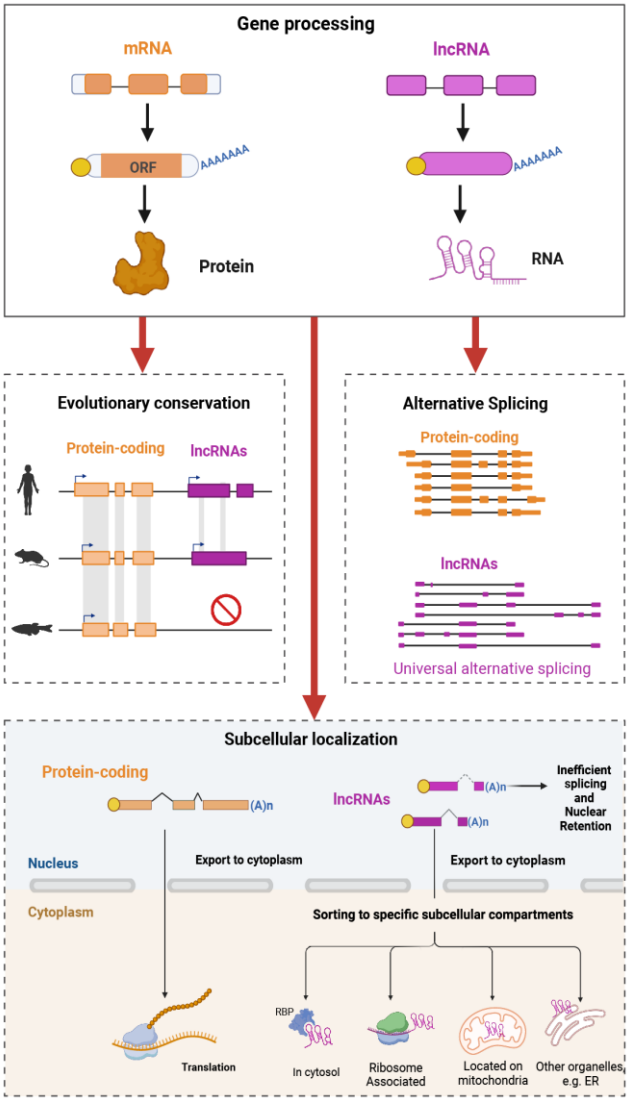


Figure 1.3. Specific features of lncRNAs compared to protein-coding genes. *Protein-coding genes and mRNAs are shown in orange, while lncRNAs are shown in magenta.*

The presence of an open reading frame in protein-coding genes allows for relatively easy identification of protein-coding sequences in the genome and enables fairly accurate assignment of their biological functions, as protein sequences can be translated back to nucleotide sequences and mapped to the genome.

Unfortunately, such an option does not exist for lncRNAs, as their sequence-structure-function relationship remains largely unknown. Recent studies have suggested that lncRNAs might have a modular architecture (Figure 1.4). This concept suggests that lncRNAs consist of distinct functional domains or modules, with each module responsible for specific aspects of their biological activity (Johnson and Guigó, 2014; Mattick et al., 2023). These domains enable interactions with various biological molecules, such as DNA, RNA, or proteins, and are connected by long linker regions that are not essential to the lncRNA's function (Johnson and Guigó, 2014; Chillón and Marcia, 2020). Alternative splicing can control the inclusion or exclusion of specific exons that contain functional domains, thereby directly shaping the functional potential of lncRNA molecules (Figure 1.4) (Mattick et al., 2023).

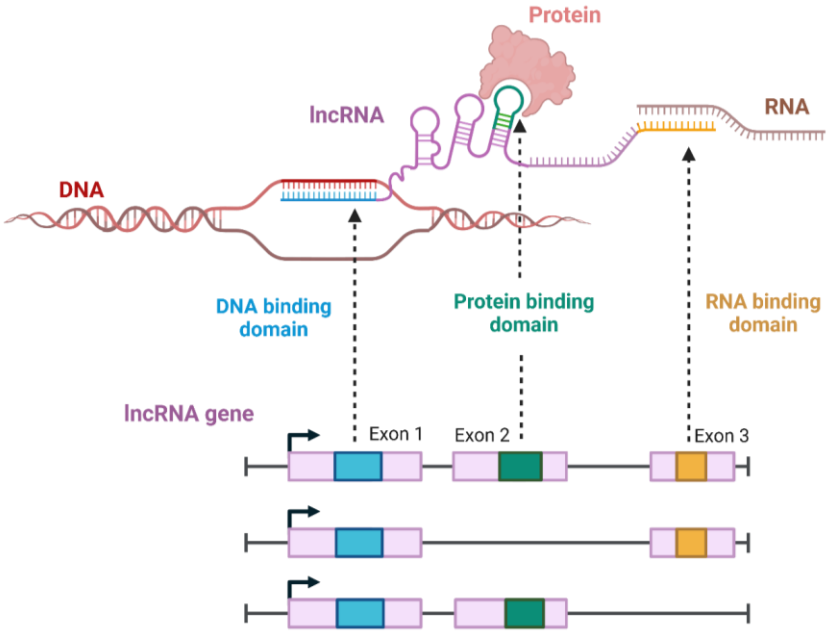


Figure 1.4. Modular architecture of lncRNAs and alternative splicing as a regulator of lncRNA functionality.

The unique features and high complexity of lncRNAs, as described above, make it challenging to identify and investigate their cellular roles. Consequently, over 97% of identified lncRNA genes remain functionally uncharacterized.

1.3. Are lncRNAs functional?

The functionality of lncRNAs is a subject of significant debate and remains one of the most actively discussed topics within the scientific community. There are two opposing perspectives on the scope of lncRNA biological functions: the conservative and the functional view. The conservative perspective argues that biological function is closely linked to evolutionary conservation at the sequence level, emphasizing protein-coding genes as the primary functional elements of the genome (Palazzo and Lee, 2015; Palazzo and Koonin, 2020). While this view acknowledges the existence of some functional lncRNAs, it regards the majority as transcriptional byproducts with little or no impact on cellular function (Palazzo and Lee, 2015; Palazzo and Koonin, 2020). In contrast, the functional view argues that lncRNAs are just as important as mRNAs, presenting the genome as a combination of both functional lncRNAs and mRNAs (Mercer et al., 2009; Mattick and Amaral, 2022; Mattick et al., 2023).

The existence of functional lncRNAs is obvious, with numerous examples with clear cellular roles and mechanisms. LncRNAs have been found to be involved in crucial biological processes including embryonic development (*H19*) (Gabory et al., 2010), X-chromosome dosage compensation (*XIST*) (Sahakyan et al., 2019) DNA repair (*LINP1*) (Zhang et al., 2016), genomic stability (*NORAD*) (Lee et al., 2016) or organogenesis (*CARMN*) (Ounzain et al., 2015). The molecular functions of lncRNAs have been found to be closely linked to their subcellular localization. Nuclear-localized long noncoding RNAs can act as transcription regulators by recruiting chromatin-modifying complexes to specific loci, organizing nuclear structures or affecting Transcription Factor binding to promoter regions (Statello et al., 2021; Bridges et al., 2021). On the other hand, cytoplasmic lncRNAs are important for signal transduction, regulation of mRNA stability and translation, or act as decoys in the sequestration of miRNAs (Noh et al., 2018; Statello et al., 2021; Bridges et al., 2021). Dysregulation of numerous lncRNAs, has also been associated with disease progression, including cardiovascular or neurological disorders (DiStefano, 2018), or has been associated with different types of cancers (*MALAT1*, *HOTAIR*) (Huarte, 2015).

When discussing lncRNA functionality, it is crucial to consider the mechanisms by which they operate. LncRNAs carry out their biological roles through five key "functional modalities" (Figure 1.5):

I. **Overlap with Regulatory DNA Elements (RNA not important for function)**: This category includes lncRNAs for which the functional element is solely a regulatory DNA region embedded within the lncRNA locus, acting as either an enhancer or a silencer, and serving a function unrelated to the transcription of the locus nor produced RNA molecule (Marchese et al., 2017; Chen et al., 2017).
II. **Transcription or Splicing Alone (RNA not important for function)**: Some lncRNAs regulate the neighboring genes by the act of transcription or splicing itself (Kornienko et al., 2013; Ali and Grote et al., 2020).
III. **Encoding Short Peptides (semi RNA-dependent function)**: A subset of lncRNAs contains short ORFs within their locus, encoding small peptides that may have significant cellular roles (Pan et al., 2022; Xiao et al., 2024).
IV. **Hosts for Small RNA Species (semi RNA-dependent function)**: LncRNAs can serve as host genes for the expression of small ncRNA species, such as miRNAs and snoRNAs (Monziani and Ulitsky, 2023).
V. **Functional lncRNA molecules (RNA-dependent function)**: These lncRNAs function through their mature RNA molecules and execute their roles via specific RNA domains within their structures, allowing interactions with DNA, proteins, or other RNA molecules. This class of lncRNAs can be further classified into cis-acting and trans-acting types, depending on whether their function is limited to the DNA region from which they were transcribed or extends to other regions (Kopp and Mendel, 2018; Statello et al., 2021; Monziani and Ulitsky, 2023).

Notably, these modes of action are not mutually exclusive, as a single lncRNA gene can carry out multiple functions through distinct mechanisms (Marchese et al., 2017; Monziani and Ulitsky, 2023). As a result, lncRNAs challenge the traditional "one gene = one function" concept typically associated with mRNAs.

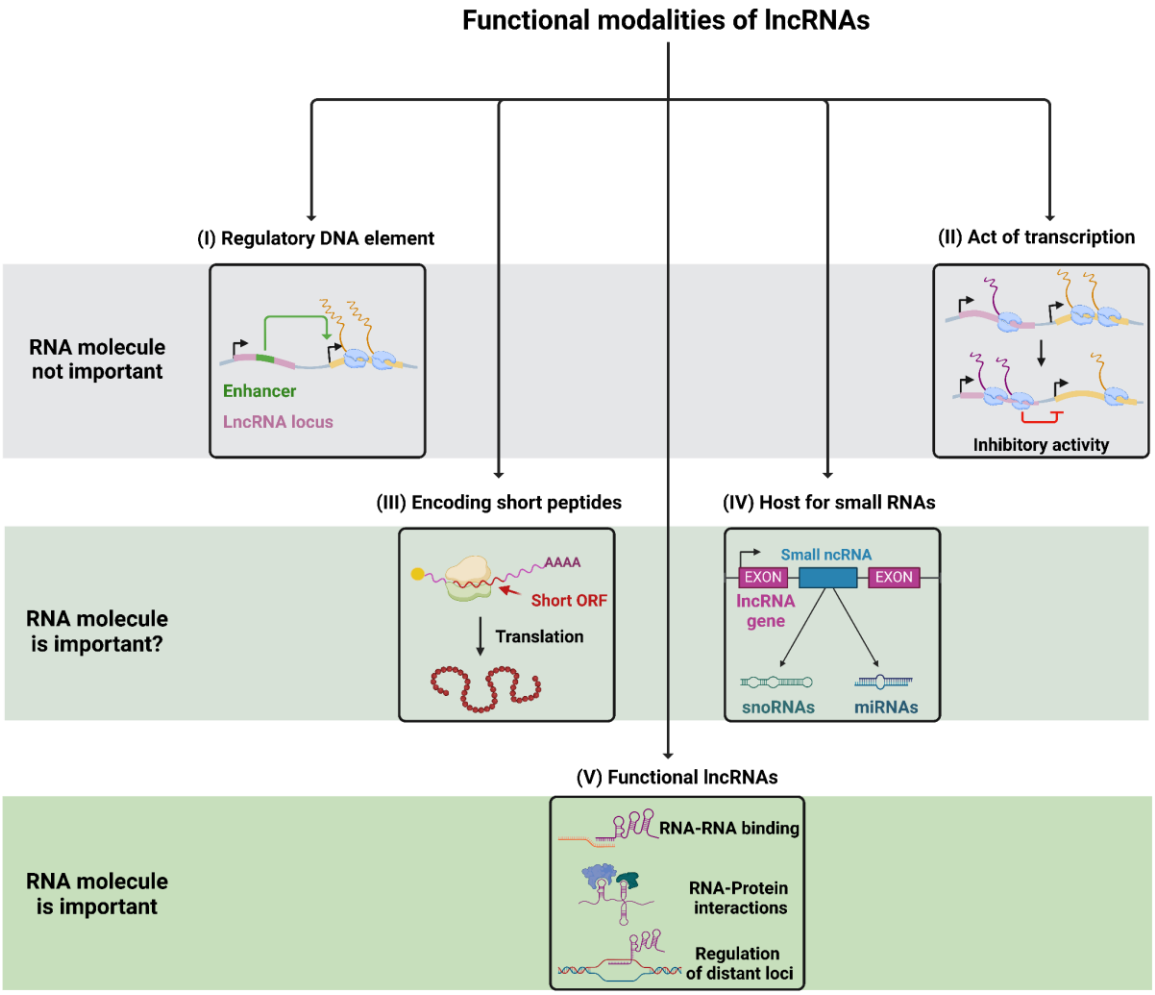


Figure 1.5. Functional modalities of lncRNAs, *either independent (grey) or dependent (green) on the mature RNA molecule.*

1.4. How do lncRNAs evolve?

LncRNAs evolve much more rapidly than protein-coding genes, and some can acquire biological functions soon after their emergence, such as lineage- or species-specific lncRNAs (Ponting et al., 2009). Recently, a mechanism has been proposed to explain the evolution of lncRNAs (Figure 1.6). According to Palazzo et al., lncRNAs initially evolve as local regulators, with their transcription providing advantages to neighboring genes. Although the RNA molecule itself may be initially dispensable for its function, it can explore the sequence space and luckily acquire sequence-specific roles over time (Palatzzo et al. 2020).

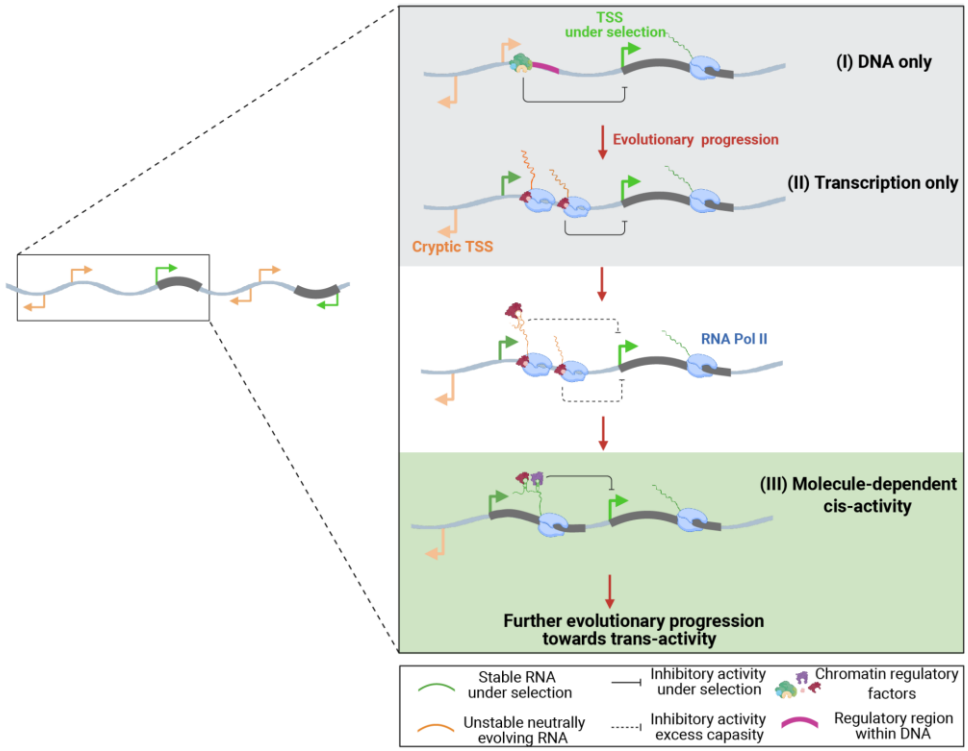


Figure 1.6. Constructive neutral evolution of lncRNA activity. *Figure based on (Palatzzo et al. 2020).*

LncRNA genes are particularly prone to mutations and genomic rearrangements, such as deletions, translocations, or endogenous retrotransposon insertions, which can introduce new functional capabilities (Yang et al., 2023; Esposito et al., 2023). The incorporation of transposable elements (TEs) plays a significant role in shaping lncRNA functions, as many TE-derived fragments serve as binding domains for RNA, DNA, or proteins, and can even influence the subcellular localization of lncRNAs (Lee et al., 2019; Fort et al., 2021). Recently, the Constructive Neutral Evolution (CNE) mechanism, which suggests that new elements emerge not de novo but through the rearrangement of pre-existing ones, has shed light on the rapid evolution of lncRNAs (Palatzzo et al. 2020). This process explains how a single lncRNA gene can acquire new functional modalities over time. This mechanism accounts also for lineage-specific lncRNAs that can gain biological functions immediately after their emergence. Furthermore, genomic modifications within non-coding regions can lead to the formation of a short Open Reading Frame (ORF), highlighting the fluidity of the boundary between lncRNAs and mRNAs. This suggests that there is no intrinsic barrier preventinga lncRNA from evolving into an mRNA over time. This highlights the need for comparative genomic studies between closely and distantly related species to fully comprehend the evolution of lncRNAs and their functional implications.

1.5. Three dimensions of lncRNA conservation

Comparative genomics has proven to be a valuable method for identifying functional regions within genomes (Perron et al., 2017; Carlevaro-Fita et al., 2019). These studies often indicate the presence of similar genes, their fragments, or even large DNA segments across various species. While such studies have greatly enhanced our understanding of protein-coding genes (Lin et al., 2011) and small non-coding RNAs, such as miRNAs (Xion et al., 2019), they are also anticipated to offer valuable insights into lncRNA genomics and functionality. However, the conservation patterns in lncRNAs may differ from those commonly observed in protein-coding genes (Diederichs, 2014; Ulitsky, 2016). Specifically, three levels of conservation have been identified for lncRNAs: (A) sequence, (B) structure, and (C) function.

A. Primary sequence conservation

The primary method for investigating evolutionary conservation of genes is nucleotide sequence homology (Figure 1.7 A). However, lncRNAs are known to evolve more rapidly than other gene classes (Hezroni et al., 2015; Ulitsky, 2016). Data from PLAR (Pipeline for LncRNA Annotation from RNA-seq data) reported that over 70% of lincRNAs (long intergenic noncoding RNAs, which do not overlap with protein-coding genes) lack sequence-similar orthologs in species separated by more than 50 million years of evolution (Hezroni et al., 2015). LncRNA sequences tend to exhibit only mild conservation across mammals and weak conservation outside mammalian species, with fewer than 100 human lincRNAs showing linear sequence conservation in the zebrafish genome (Ulitsky et al., 2011). This can be exemplified by *MALAT1* (*Metastasis Associated Lung Adenocarcinoma Transcript 1*), one of the most highly conserved lncRNAs (Weghorst el al., 2024), whose sequence is well preserved between humans, chimpanzees, and mice, but shows only partial conservation level between humans and zebrafish (Figure 1.7 B). Moreover, it was shown that conserved regions are typically limited to short segments of 50-300 nucleotides, surrounded by rapidly evolving sequences (Hezroni et al., 2015; Ulitsky, 2016). As a result, many lncRNA homologs may be underestimated because the tools for comparative analysis were designed for protein-coding genes and expect long, contiguous sequences, which is unlikely due to the modular architecture of lncRNAs.

Consequently, new approaches are needed to detect these short stretches of conservation. Two methods, SEEKR (Kirk et al., 2018) and LncLOOM (Ross et al., 2021), have been developed to analyze micro-stretches of homology scattered throughout lncRNA transcripts, revealing that many of them correspond to DNA, protein, or miRNA-binding motifs, which aids in the functional classification of lncRNAs (Constanty and Shkumatava, 2021).

B. Structure conservation

Despite differences at the primary sequence level, many lncRNA molecules demonstrate structural conservation (Figure 1.7 A). For these transcripts, preserving base-pairing properties is more important than maintaining the exact nucleotide sequence (Diederichs, 2014). As a result, reciprocal mutations can still sustain the same secondary structure. This is exemplified by the *COOLAIR* gene, an lncRNA that plays a crucial role in regulating the major plant developmental gene *FLOWERING LOCUS C (FLC)* (Jiao et al., 2019). *COOLAIR*, despite relatively low nucleotide sequence identity, exhibits remarkable evolutionary conservation of its structures across plant species (Hawkes et al., 2016). Moreover, the *CYRANO* lncRNA, responsible for morphogenesis and neurogenesis in zebrafish (Ulitsky et al., 2011) has preserved its cloverleaf structure for over 400 million years, which is crucial for its specific interactions with RNA-binding proteins (Jones et al., 2020).

C. Functional conservation

The final class includes lncRNAs that, despite lacking conserved sequence or structure, retain functional roles (Figure 1.7 A). A key example is sex chromosome dosage compensation, regulated by distinct lncRNAs across species: *roX1/roX2* in Drosophila (Franke and Baker, 1999), *Rsx* in opossum (Grant et al., 2012), and *Xist/XIST* in mammals (Hong et al., 2000). However, the marsupial lncRNA *Rsx* silences the X chromosome in opossums, functioning similarly to *Xist* without identifiable sequence homology (Grant et al., 2012; Karner et al., 2020). *Lnx3* occupies the *Xist* locus without affecting dosage compensation (Karner et al., 2020). Another example is human *JPX* and its mouse homolog, which, despite significant divergence in nucleotide sequences and RNA secondary structures, preserve their binding to the CTCF protein and maintain a conserved role in X chromosome inactivation (Karner et al., 2020).

1.6. Positional conservation as the fourth dimension of lncRNA conservation

Given the rapid evolution of lncRNAs, identifying homologous genes cannot rely exclusively on sequence and structural homology, requiring alternative approaches that do not depend on these criteria. One promising approach is positional conservation, which identifies potentially functional lncRNAs by analyzing conserved genetic blocks across various species (Figure 1.7 A) (Bryzghalov et al., 2021; Szcześniak et al. 2021). This strategy of searching for syntenic regions has effectively revealed lncRNA orthologues, even among distantly related organisms (Huang et al., 2024). By utilizing positional conservation, researchers have identified hundreds to thousands of lncRNAs with conserved positions across vertebrate genomes, many of which were overlooked by sequence alignment-based methods (Ulitsky et al., 2011; Ulitsky, 2016, Hezroni et al., 2015).

Interestingly, there are increasing reports supporting the conserved function of lncRNA syntologs between humans and zebrafish across 450 million years of evolution. For example, Huang et al. identified 570 lncRNAs with conserved genomic locations between humans and zebrafish (coPARSE-lncRNAs) (Huang et al., 2024). Although fewer than 3% of these coPARSE-lncRNAs (17 lncRNAs) showed detectable sequence similarity, many of them appear to have conserved lncRNA-interacting proteomes with subsequent preserved function (Huang et al., 2024). Moreover, Sabaté-Cadenas et al. identified *casc15* positionally conserved lncRNA to be involved in the regulation of melanogenesis (Figure 1.7 C). This syntelogous lncRNA appears to display conserved function between humans and zebrafish despite no detectable sequence similarity (Sabaté-Cadenas et al., 2024). This was demonstrated by rescue experiments in which the expression of human *CASC15* in zebrafish restored *casc15* function and reduced melanoma formation. The authors also revealed that the conserved function of *CASC15* is mediated by the preservation of RNA-binding protein partners that interact with both zebrafish and human *CASC15* transcripts (Sabaté-Cadenas et al., 2024).

Altogether, these reports show that the conserved function of lncRNAs between distant species, such as humans and zebrafish, is often linked to the preservation of genomic position and the conservation of RNA-binding protein (RBP) interactions (Huang et al., 2024; Sabaté-Cadenas et al., 2024). Thus, positional conservation offers a valuable opportunity to identify potentially functional lncRNAs in distant species without relying on sequence and structural homology.

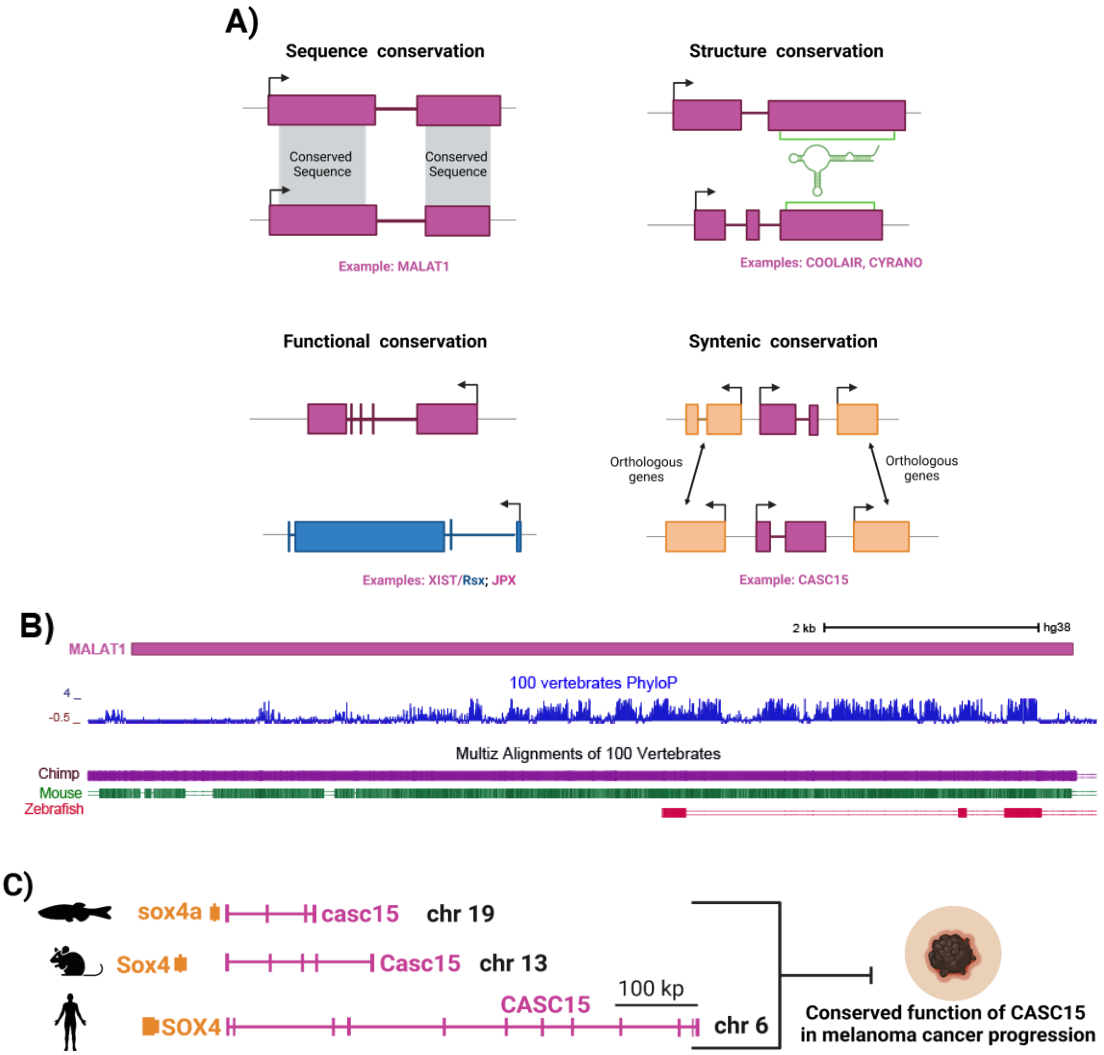


Figure 1.7. A*)* Four dimensions of lncRNA evolutionary conservation. *B)* *Sequence conservation of the MALAT1 lncRNA across human (pink), chimp (purple), mouse (green), and zebrafish (red).* ***C)*** *Syntenic conservation of CASC15 locus across human, mouse and zebrafish.*

1.7. Zebrafish as compelling animal model

Over the years, animal models have played a crucial role in exploring fundamental physiological and pathological processes. The choice of a model organism for studying biological mechanisms is not arbitrary and requires careful consideration. This is especially true in the case of lncRNAs, where our understanding solely depends on loss-of-function experiments.

Zebrafish (*Danio rerio*) - a small freshwater fish native to Southeast Asia has emerged as an attractive, fully-developed organism that is poised to significantly improve our understanding of vertebrate biology (Rahman Khan and Sulaiman Alhewairini, 2018).

Zebrafish shares the same major organs and tissues with the human body, therefore it has a very similar anatomy and physiology to ours (Figure 1.8) (Rahman Khan and Sulaiman Alhewairini, 2018). Moreover, comparative genomic studies revealed that at least 70% of human genes have a zebrafish counterpart (Howe, 2013) thus showing remarkable similarity in their genomes. Those high genomic and physiological similarities ensure that information acquired through zebrafish is more accurate than obtained by *in vitro* studies or in non-vertebrate organisms and can be easily extrapolated to human biology.

Importantly, zebrafish also provides significant advantages over widely used rodents (mice or rats) that make it a competing model for biomedical research. These include easier and lower-cost maintenance and higher fertility rate. While female mice manage to breed only multiple times per year, with litter sizes up to a dozen pups (Perlman, 2016), Danio rerio can produce hundreds of eggs at weekly intervals. What is even more, *ex utero* fertilization and development of zebrafish embryos allow for detailed and easy exploration of embryogenesis from fertilization through cell division, migration, differentiation, and organogenesis (Kimmel, 1995; Machikhin et al., 2020).

These aspects make zebrafish a perfect (and unfortunately still undervalued) model to study vertebrate development and molecular genetics. Indeed, zebrafish have greatly contributed to a better understanding of basic physiological and pathological processes (Choi et al., 2021). It has also helped to uncover the functions of many protein-coding genes (Lieschke and Currie, 2007). Taking into consideration that a subset of lncRNAs are evolutionarily conserved across vertebrates (Ulitsky et al., 2011; Huang et al., 2024), zebrafish appear as a promising and competent model system that could be used to study lncRNA evolution and function *in vivo*.

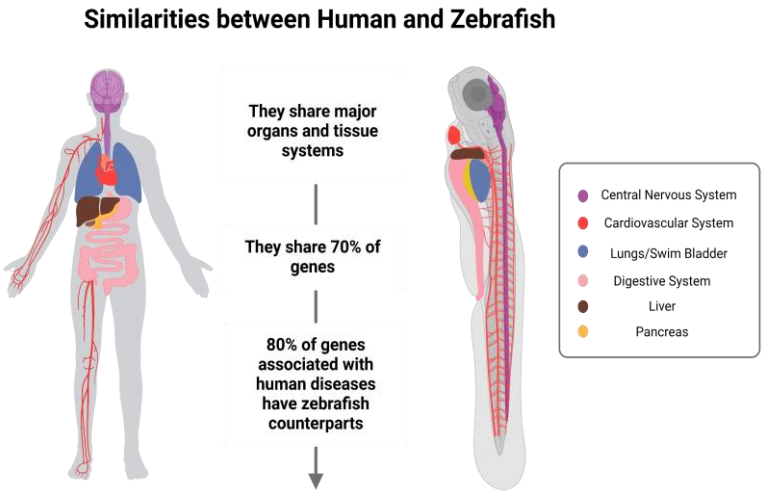


Figure 1.8. Similarities between human and zebrafish.

1.8. Zebrafish lncRNA annotations

Accurate annotations are essential across all areas of biological and biomedical science, as they provide a detailed map of gene locations, the functional units of the genome. Annotations for both protein-coding and lncRNA genes have a hierarchical structure (Figure 1.9). They consist of gene loci, each made up of one or more overlapping transcripts, which in turn are composed of one or multiple exons. Comprehensive annotation involves identifying every locus, transcript, and exon expressed at any time point and in any cell type throughout an organism's entire lifespan. While the annotation of human protein-coding genes is nearly complete (Amaral et al., 2023), lncRNA annotations remain largely incomplete due to their unique features. Unlike protein-coding genes, lncRNAs are challenging to annotate due to their weak expression, high tissue specificity, and less understood sequence-function relationships (Uszczyńska-Ratajczak, 2018; Amaral et al., 2023). Additionally, weak evolutionary conservation of lncRNA sequences makes the identification of their orthologs or paralogues through sequence similarity challenging. As a result, lncRNA annotation relies almost exclusively on direct transcriptomic evidence.

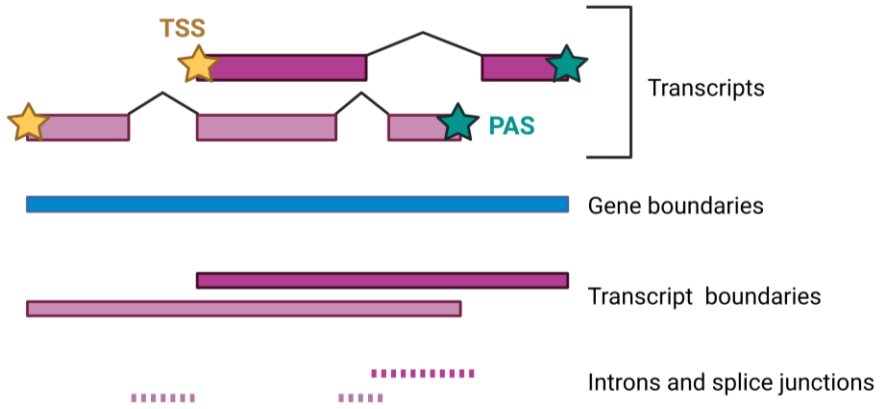


Figure 1.9. The key structural elements of lncRNA loci to be annotated. *Annotations consist of gene loci (depicted in blue), each containing one or more partially overlapping transcripts (depicted in pink), which are further composed of one or more exons (represented by pink rectangles). The Transcription Start Site (TSS) is represented by a yellow star, while the Polyadenylation Signal is indicated by a green star.*

Although human lncRNA annotations are incomplete, the issue is particularly significant in zebrafish. Despite being the third-best annotated and well-characterized genome, its annotation quality still lags behind that of humans and mice, which limits zebrafish utility in lncRNA function-specific research.

Different genome annotation approaches

Genome annotations can be created using two primary approaches: (1) manual curation and (2) automated annotation through transcriptome assembly, with each method possessing its own strengths and limitations (Uszczyńska-Ratajczak, 2018). The manual annotation follows rigorous, standardized procedures that combine genomic and transcriptomic evidence, often supplemented with experimental validation. While this approach produces highly accurate annotations, it advances at a slower pace (Harrow et al., 2006; McDonnell et al., 2018). Only a small fraction of zebrafish genes and transcript models have been manually curated, primarily by the Human and Vertebrate Analysis and Annotation (HAVANA) group under the Vertebrate Genome Annotation (VEGA) project. These efforts aimed to produce manual gene annotations for multiple species, including zebrafish, and integrate them into Ensembl (Loveland, 2005; Wilming et al., 2008). Unfortunately, manual curation of the zebrafish genome by HAVANA group was discontinued in 2016 (last release VEGA67). As a result, recent zebrafish transcript models are primarily based on automated strategies.

The automated annotation is mainly based on application of RNA sequencing (RNA-seq), which in recent years enhanced the detection of lncRNAs across various species. In these studies, short RNA-seq reads are mapped to the reference genome, followed by either assembling transcript models or performing de novo assembly before aligning them to the genome. Several filters are applied to annotate lncRNAs, including a minimum transcript length of over 200 bp, and the assessment of coding potential using tools like the Coding Potential Calculator (CPC2) (Kang et al., 2017), Coding Potential Assessment Tool (CPAT) (Wang et al., 2013), and PhyloCSF (Lin et al., 2011). Transcripts longer than 200 bp with no detectable coding potential are classified as putative lncRNAs and included in publicly available lncRNA databases. Automated annotation is favored for its time and cost efficiency. However, it often results in incomplete transcript models, including missing terminal exons and other artifacts (Lagarde et al., 2017; Uszczyńska-Ratajczak et al.,2018). Moreover, the zebrafish genome is approximately half the size of most mammalian genomes, yet its transcriptome complexity is comparable to that of humans and mice (Howe et al., 2013; Shehwana and Konu, 2019). Historically, zebrafish gene models were primarily constructed using relatively shallow short-read RNA-seq data (50 vs. 250 million reads per sample for human in the ENCODE project), mainly derived from early developmental stages (Pauli et al., 2012; Wang et al., 2017).

Given the transcriptomic complexity of zebrafish, creating accurate gene models with such limited data is nearly impossible. As a result, zebrafish lncRNA annotations lag significantly behind those of humans and mice in terms of accuracy and completeness.

Zebrafish lncRNA annotation bias

Another significant challenge in using zebrafish for biomedical studies and modeling human diseases is the bias in genome annotations. Modern human annotations are predominantly focused on adult tissues and cell lines due to ethical considerations (Frankish et al., 2023; Lorenzi et al. 2021). In contrast, zebrafish genome annotations are primarily oriented towards developmental stages, reflecting its extensive use in developmental research (Vesterlund et al. 2011; White et al., 2017; Pauli et al., 2012; Zhou et al., 2024). Given that lncRNAs exhibit highly spatio-temporal-specific expression patterns, performing a comparative analysis with such biased data is nearly impossible and would likely result in numerous incorrect conclusions.

Ecosystem of zebrafish lncRNA annotations

Overall, there are **six** major lncRNA annotations publicly available for zebrafish. Collectively, these databases report thousands of lncRNAs identified across various tissues and developmental stages.

(1) Ensembl (Harrison et al., 2024) and **(2) RefSeq** by NCBI (O'Leary et al., 2016) are two reference annotations for zebrafish genome that encompass both coding and non-coding gene models, which are either manually curated or automatically assembled.

(3) NONCODE (Zhao et al., 2021) is one of the earliest databases that integrates annotations from a mix of manual literature searches and other sources. It catalogs approximately 4,852 lncRNA transcript isoforms from 3,503 lncRNA genes in zebrafish.

(4) zflncRNApedia (Dhiman et al., 2015) a zebrafish-specific resource based on data integration and manual curation of literature lncRNAs, it compiles 2,267 lncRNAs from key zebrafish studies (Ulitsky et al., 2011; Pauli et al., 2012; Kaushik et al. 2013).

(5) ZFLNC (Hu et al., 2018) currently the largest lncRNA database for zebrafish. It extensively integrated lncRNA data from NCBI, Ensembl, NONCODE, zflncRNApedia, literature, and independent RNA-seq analyses, identifying 13,604 lncRNA genes and 21,128 lncRNA transcripts.

(6) LncRBase (Das et al., 2020) provides transcriptomic information for 13,866 lncRNA transcripts and also integrates the data on predicted subcellular localization, co-localized miRNAs, and tissue-specific expression for these lncRNAs.

The lncRNA annotation quality

Despite significant advances in zebrafish lncRNA annotations, current resources remain incomplete in two key ways: (1) many annotated gene models are partial, representing only fragments of full gene structures with inaccurately assessed transcription start and termination sites, and (2) some loci may be entirely absent from the annotations.

A comparison of the Ensembl gene set for zebrafish (v104) (Harrison et al., 2024) with the GENCODE sets for human (v47) and mouse (vM36) (Kaur et al., 2024), which are the most comparable annotation sets, revealed that while all species have a similar number of protein-coding genes, the total number of annotated lncRNAs in zebrafish is more than 10 times lower than in the human and mouse genomes (Figure 1.10).

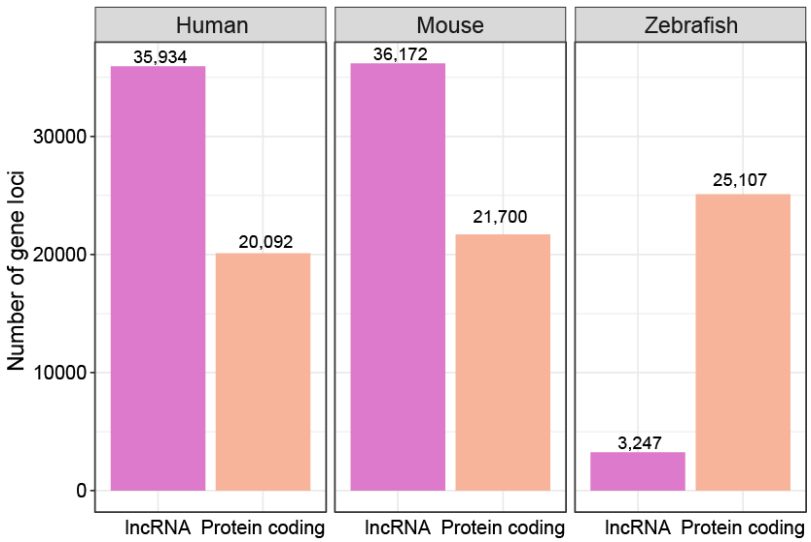


Figure 1.10. Comparison of the number of gene loci *(protein-coding genes in orange and lncRNA genes in purple) annotated in human (GENCODE v47), mouse (GENCODE* vM36*), and zebrafish (Ensembl v104).*

What is more concerning is that zebrafish lncRNA resources have not been updated in several years. In contrast, over the past seven years, ~20,000 new lncRNA loci have been identified in humans and ~23,000 in mice, while only a few new loci have been detected in zebrafish (Figure 1.11).

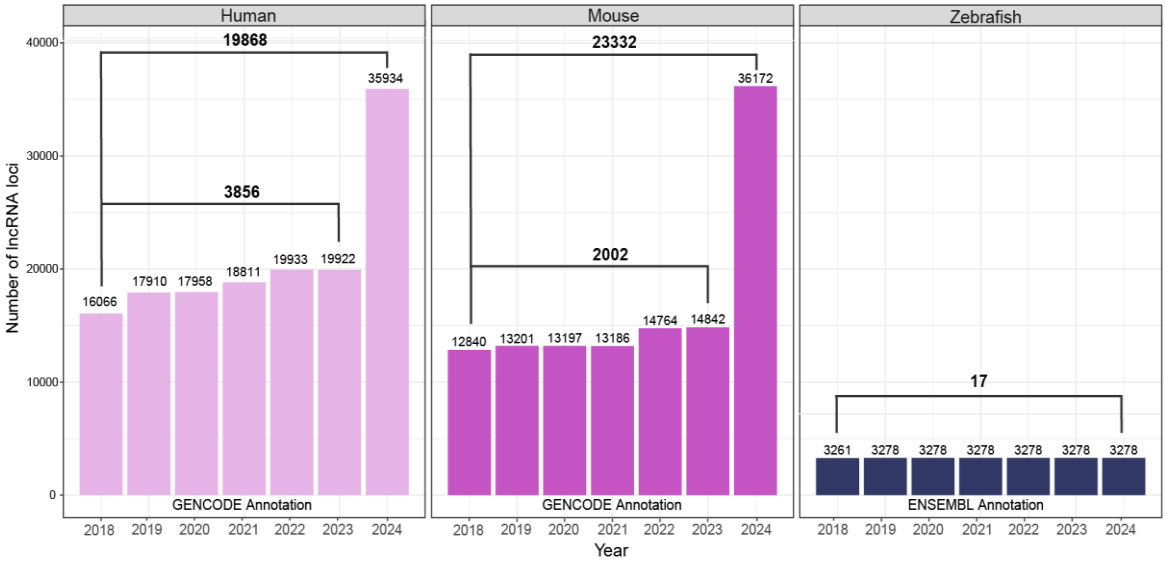


Figure 1.11. Updates to lncRNA gene catalogs over the years. *Updates to human lncRNA catalogs are shown in light pink, mouse updates are illustrated in bright pink, and updates for the zebrafish genome are depicted in navy blue.*

Detailed analyses revealed that all zebrafish lncRNA resources exhibit poor performance in terms of completeness, with fewer than 30% of lncRNA models being full-length, with RefSeq (9%) and Ensembl (20%) being the most affected. Furthermore, zebrafish lncRNA catalogs, on average, identified fewer than two transcript isoforms per gene locus, compared to six isoforms in the human lncRNA catalog (Figure 1.12).

Moreover, existing annotation methods and databases show limited consensus as significant discrepancies were identified between reference gene catalogs for the zebrafish genome, such as those provided by NCBI's RefSeq and EMBL-EBI's Ensembl (Lawson et al., 2020). Analysis by Lawson et al. revealed that thousands of gene models were missing from each resource. Notably, Ensembl and RefSeq also represented alternative annotations for 3'-UTRs (Lawson et al., 2020). Similar low overlap has been observed in human gene catalogs, leading to the initiation of the MANE project (Matched Annotation from NCBI and EMBL-EBI). This project aims to harmonize human gene and transcript annotations from RefSeq and Ensembl and to establish a genome-wide set of representative transcripts (Morales et al., 2022).

Unfortunately, there is no equivalent initiative for the zebrafish genome. Low consensus is observed not only in gene loci and transcript structures but also in lncRNA nomenclature. The absence of official gene names for many lncRNAs complicates efforts to track and consolidate existing knowledge about specific lncRNAs.

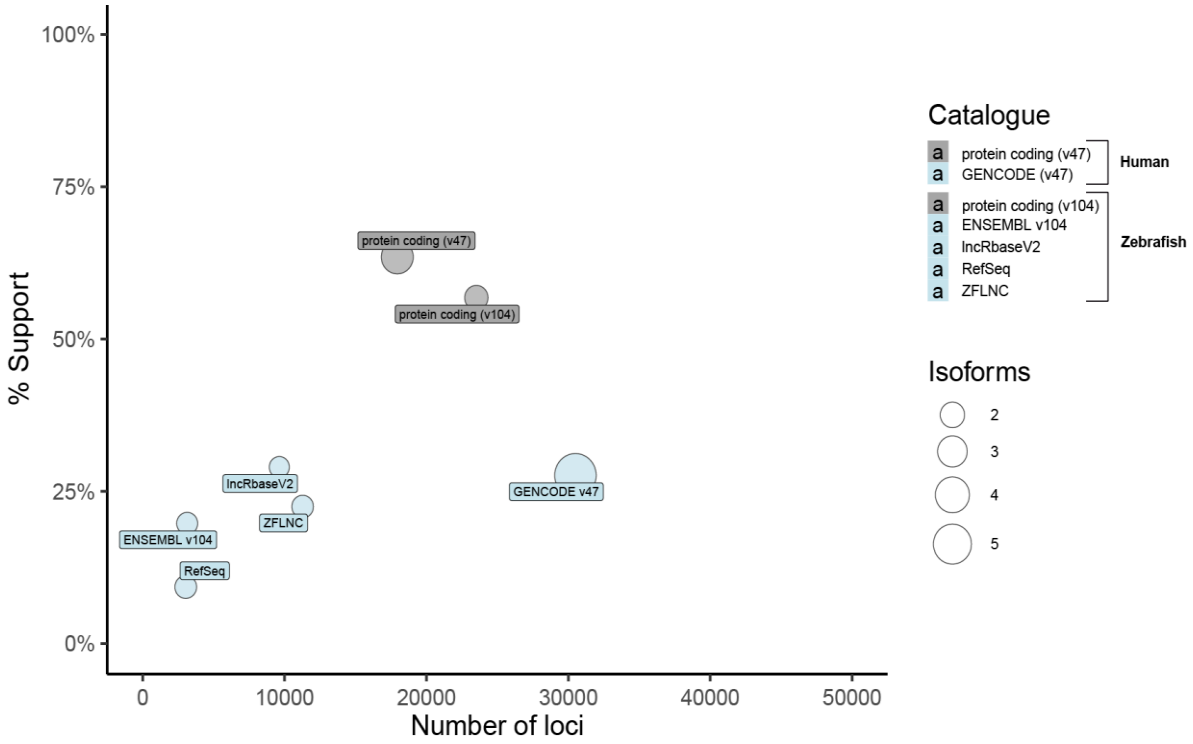


Figure 1.12. Comparison of quality metrics across zebrafish lncRNA annotations. *If the genome build of analyzed annotation predates GRCz11, it was lifted to this version. To ensure consistency, assembly patches were excluded, and gene loci boundaries were redefined using buildLoci (https://github.com/julienlag/buildLoci), resulting in expected differences in gene locus counts compared to the reported lncRNA in each annotation. The x-axis represents the total number of gene loci in each annotation, while the y-axis indicates the percentage of transcript structures supported by a CAGE cluster (DANIO-CODE) and polyA signal. The diameters of the circles represent the mean number of transcripts per gene. The "protein-coding" and "GENCODE" (v47) categories represent confidently annotated GENCODE protein-coding and lncRNA genes for the human genome, respectively (Uszczyńska-Ratajczak, 2018, Kaur et al., 2024). Protein-coding catalogs are shown in gray, while lncRNA catalogs are shown in blue.*

Altogether, this demonstrates that zebrafish lncRNA annotations are of poor quality. Therefore, it is clear that effective validation of lncRNA functions in zebrafish will require significant advancements in its genome annotation. Completing the zebrafish lncRNA catalog will necessitate the development of innovative technologies to enhance the completeness of transcript models and aid in the identification of novel lncRNA loci.

1.9. Towards comprehensive and complete lncRNA annotation in zebrafish

In recent years, the scientific community has placed significant emphasis on enhancing gene catalogs for various organisms, including zebrafish, as exemplified by initiatives such as The Earth BioGenome Project (Lewin et al., 2018; Lewin et al., 2022) or Darwin Tree of Life Project (https://wellcomeopenresearch.org/treeoflife). These efforts are focused on developing innovative, time- and cost-efficient methods capable of producing complete and accurate genome annotations that match the quality of manual curation, but with minimal human involvement. Recent advancements in key technologies for targeting and sequencing full-length lncRNA molecules hold the potential to directly tackle the two main challenges in lncRNA annotation: low target abundance and incomplete transcript models (Lagarde et al., 2017; Uszczyńska-Ratajczak et al.,2018).

The first initiative is to increase the use of long-read sequencing, or Third-Generation Sequencing (TGS), for gene annotation (Lagarde et al., 2017; Pardo-Palacios et al., 2024). Recently developed technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), produce extremely long reads (>10 Kb) that span entire RNA molecules, allowing direct investigation of alternatively spliced transcripts and offering great potential for exploring transcriptome variation in depth (Nudelman et al., 2018; Uszczyńska-Ratajczak et al.,2018; Paoli-Iseppi et al., 2021; Wohlers et al., 2013). Although long-read sequencing platforms resolve the issue of transcriptome assembly, their sequencing depth is often insufficient to cover lowly expressed transcripts, such as lncRNAs. Moreover, while third-generation sequencing (TGS) has been available for a decade, there are only few zebrafish long-read RNA-seq experiments, primarily focused on very specific biological contexts (Nudelman et al., 2018; Mehjabin et al., 2019). Therefore, expanding its use is expected to greatly enhance zebrafish gene annotations and benefit further experimental analyses.

Although there have been significant technological advances in long-read RNA sequencing in recent years, accurate end-to-end profiling of RNA transcripts remains challenging (Carbonell-Sala et al. 2024), primarily due to technical limitations in RNA-seq library preparation. Most commercially available library preparation kits often employ SMART (Switching Mechanism At RNA Termini) technology (Figure 1.13), a simple and widely adopted cDNA synthesis approach that does not inherently

ensure 5’-to-3’ completeness, often resulting in cDNA 5’-ends that fall short of the actual Transcription Start Sites (TSSs) (Lagarde et al., 2017; Carbonell-Sala et al., 2023). Accurate identification of lncRNA transcription start sites (TSSs) is crucial for knock-out loss-of-function studies, where precise targeting of Cas9 molecules to gene promoters is essential. These studies should focus on transcripts with well-defined 5′ ends. Another limitation of the TSO method is that standard selection of polyadenylated transcripts using oligo(dT) primers alone is often insufficient in removing rRNA molecules, necessitating additional, costly rRNA depletion steps (Carbonell-Sala et al., 2023).

So far, several in-house-developed approaches have been established to increase 5'-end completeness and most of them are focused on specific selection of 5’-capped transcripts. Selecting the 5'-capped fraction of the transcriptome also has the advantage of efficiently removing rRNAs—highly abundant RNA molecules that can reduce the depth of RNA sequencing. First method called Cap0-seq employs RNA-binding properties of the human IFIT1, an antiviral protein recognizing cap 0 RNAs, to pull-down 5'-capped RNA molecules (Nowacka et al., 2022). Unfortunately, this method was primarily optimized for short-read Illumina sequencing, which would require further adaptation for long-read platforms. The alternative method, 5TERAseq uses several enzymatic treatments and subsequent 5’-adapter ligation to specifically select 5’-capped transcripts (Ibrahim et al., 2021). This method has a few potential drawbacks. First, this method was specifically optimized for direct RNA ONT sequencing, which suffers from lower sequencing yield and accuracy compared to Nanopore cDNA sequencing protocols (Grünberger et al., 2021; Carbonell-Sala et al. 2024). Moreover, adapter ligation to RNA molecules often has low efficiency, necessitating high RNA input to achieve a sufficient sequencing library amount. Additionally, T4 RNA ligases may show preferential bias towards structural features within RNA molecules and adapters, potentially leading to non-uniform transcriptome coverage (Zhuang et al. 2012). Recently CapTrap-seq has emerged as a highly competitive method (Carbonell-Sala et al. 2024). It employs two consecutive rounds of selection to specifically enrich full-length transcripts in cDNA libraries (Figure 1.13). In the first round, a CAP trapper technique, involving the chemical introduction of biotin into the diol residue of cap structures, is used to selectively capture 5'-capped molecules. This step not only addresses the issue of 5'-end incompleteness but also efficiently removes rRNA molecules. The second round involves a double-stranded linker ligation step to single-stranded

cDNA (sscDNA), a highly specific reaction that accurately recognizes cap and poly(A) tail structures while protecting cDNA molecules from degradation. CapTrap-seq is a versatile method validated on both ONT and PacBio sequencing platforms, demonstrating efficiency with human and mouse samples (Carbonell-Sala et al. 2024). It enables reliable transcript reconstruction and produces a significantly high proportion of full-length transcripts. CapTrap-seq is currently used to generate transcriptome data for the GENCODE project, with numerous CapTrap-seq transcript models already included in the GENCODE gene set. One of its limitations is its complex, multi-stage procedure may result in detecting shorter RNA molecules compared to simpler cDNA preparation methods like Template Switching Oligo (TSO) (Carbonell-Sala et al. 2024).

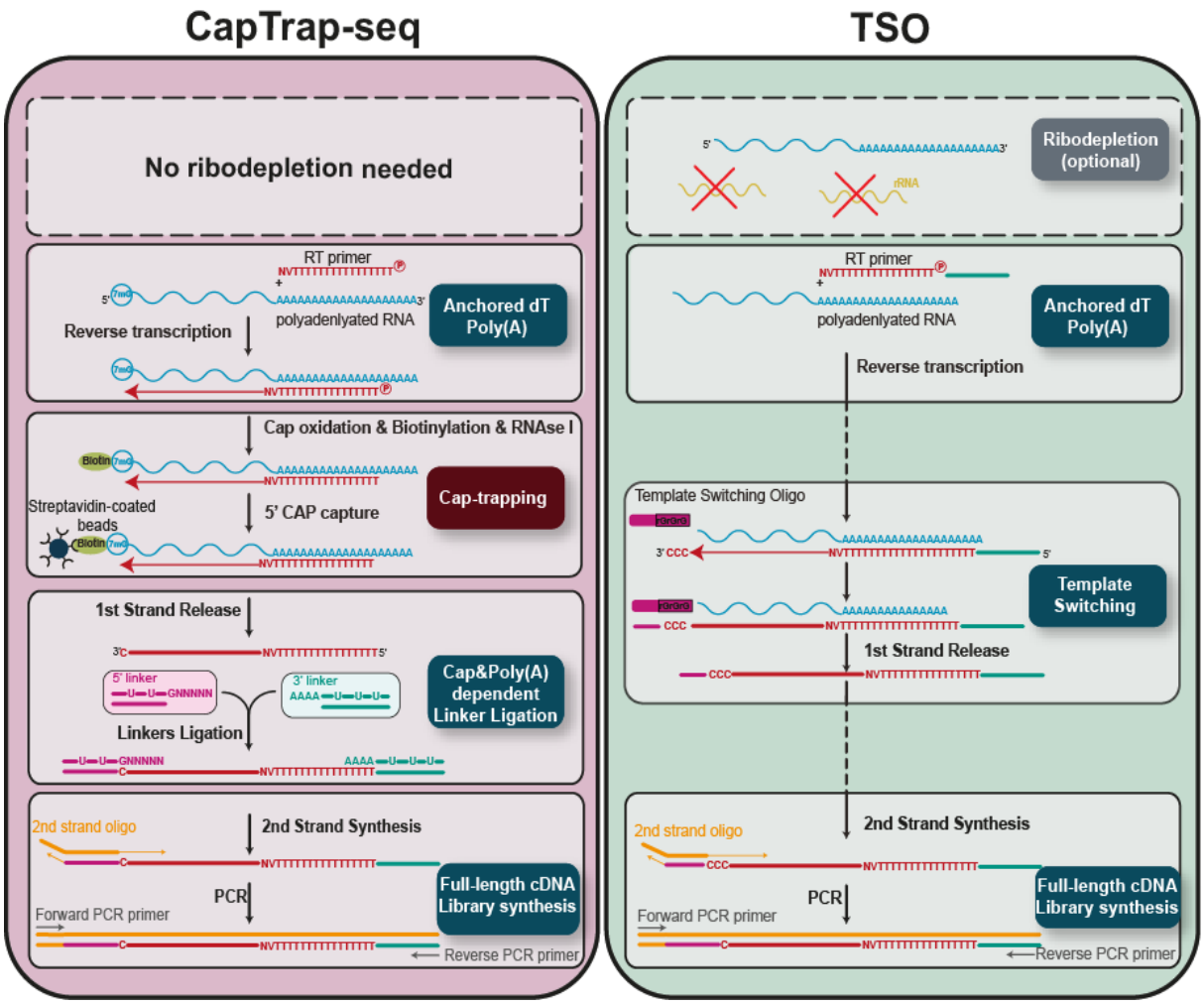


Figure 1.13. Comparison of CapTrap-seq and TSO full-length library preparation protocols.

LncRNAs are typically expressed at significantly lower levels than protein-coding genes, accounting for approximately 5% of the polyadenylated fraction of the transcriptome (Carbonell-Sala et al., 2023). This poses a significant challenge for annotation, as lncRNAs are likely to be underrepresented at the depth offered by long-read sequencing. Specialized methods are therefore needed to selectively enrich lncRNAs in sequencing libraries. RNA CaptureSeq addresses this issue by employing oligonucleotide probes to enrich lncRNA sequences, thereby increasing their representation to over 25% and significantly enhancing coverage (Lagarde et al., 2017). This method has proven effective in human and mouse tissues, enhancing sensitivity and revealing novel transcripts and gene loci that are often missed with conventional sequencing methods. Developed by the GENCODE Consortium, Capture Long Sequencing (CLS) is an advancement of CaptureSeq (Lagarde et al., 2017; Carbonell Sala et al., 2021). CLS integrates enhanced sequencing coverage from cDNA capture with the increased accuracy of transcript models provided by long reads (Figure 1.14). Capture probe libraries can be custom-designed to target known lncRNAs, refining existing annotations and exploring unknown lncRNAs in suspected regions. However, emphasis should be placed on identifying and improving the annotation of biologically relevant lncRNAs, particularly those that are positionally conserved, to facilitate their functional characterization.

With effective methods for full-length library preparation, long-read sequencing, and RNA capture now available, we are well-positioned to advance the study of low-expressed transcripts, particularly lncRNAs. Notably, the combination of CapTrap-seq with the CLS approach has already demonstrated significant effectiveness in improving lncRNA annotation for the human and mouse genomes (Kaur et al., 2024). This progress paves the way for achieving a comprehensive lncRNA annotation, ultimately providing detailed maps of the entire lncRNA landscape expressed throughout the lifespan of any organism, including zebrafish.

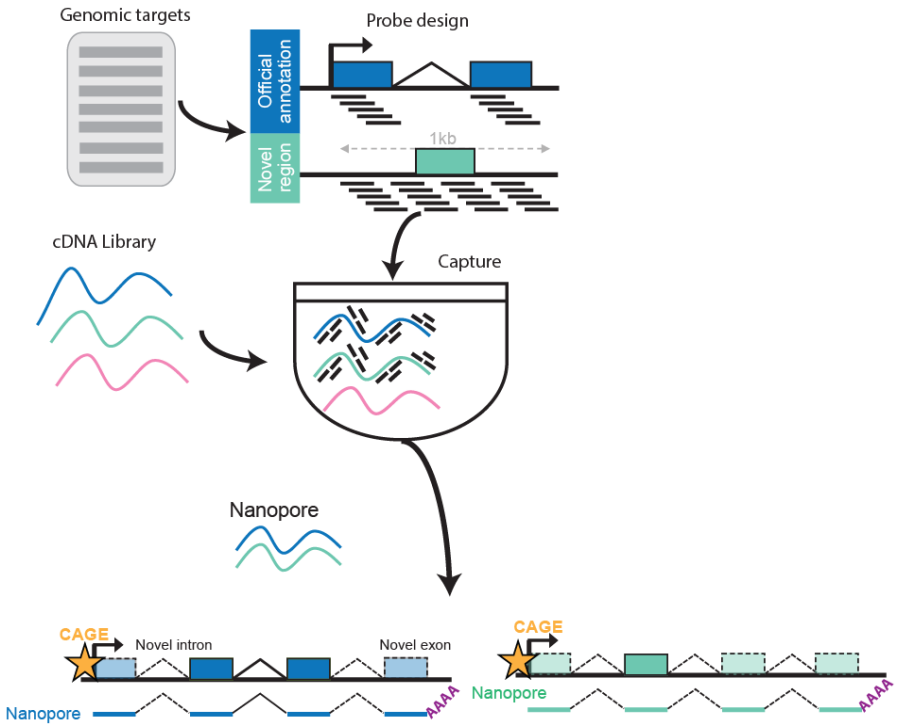


Figure 1.14. ***The workflow for automated, high-quality transcriptome annotation using CapTrap-CLS.*** *CapTrap-CLS can be used to improve existing annotations (blue) or to extend them (turquoise). Capture probes are designed to tile across targeted regions of the genome. Long-read sequencing libraries are prepared from the enriched samples. The long-read sequencing data is then used to construct transcript models. The completeness of each transcript is evaluated by its proximity to CAGE clusters (indicated in orange star) and transcription termination sites, which are identified through non-genomically encoded polyA sequences in Nanopore reads. Novel exons are represented with rectangles that have lighter shading and dashed outlines, indicating regions that were not previously annotated or are newly discovered.*

2. OBJECTIVES

Project motivation

Although thousands of lncRNA loci have been identified across diverse organisms, the vast majority (~97%) remain functionally uncharacterized. Accurate and comprehensive transcriptome annotations are crucial for assigning functions to lncRNAs, yet these RNAs are generally annotated with less precision and completeness compared to protein-coding genes. Most of the current functional insights come from loss-of-function studies in human cell lines and mouse models, limiting our understanding of their biological roles and functional evolution. To fully understand lncRNA functions, it is crucial to investigate them across a broader range of species, especially beyond mammals.

Zebrafish has recently emerged as a promising vertebrate model for advancing our understanding of lncRNA functionality. However, its potential is constrained by the quality of current lncRNA annotations, which lag behind those of humans and mice in both the number of identified gene loci and the accuracy of gene models. Several factors contribute to these limitations: a genome annotation bias toward developmental stages, incomplete gene models due to reliance on short-read sequencing, and underuse of more accurate long-read methods. Furthermore, shallow RNA sequencing lacks the sensitivity to detect lowly expressed lncRNAs. These challenges create significant gaps in the functional characterization of zebrafish lncRNAs, limiting insights into their biological roles and evolution.

The objectives of my PhD project were structured to overcome specific challenges in zebrafish lncRNA annotation as follows:

1. **Identify orthologous lncRNAs:** Employ a synteny-based approach to detect orthologous lncRNAs in humans and mice in a sequence-independent manner. This method focuses on evolutionarily and positionally conserved lncRNAs, which are likely to be functional, with the aim of improving their annotation and supporting their functional characterization.

2. **Eliminate annotation bias:** Focus transcriptomic analysis on complex and biologically relevant adult zebrafish organs with human counterparts to reduce the emphasis on developmental stages in the zebrafish genome annotation.

3. **Improve lncRNA annotation completeness:** Develop and refine a CapTrap-seq full-length library preparation method utilizing long-read Oxford Nanopore Sequencing to generate high-quality libraries with accurately defined transcription start sites (TSS), thereby enabling precise downstream experimental design.

4. **Enhance representation of lowly expressed lncRNAs:** Integrate the CapTrap-seq full-length library preparation method with a cDNA capture step, designing capture probes to improve existing gene models and identify novel lncRNAs. Target regions include small RNAs, enhancer elements, and ultraconserved noncoding elements (UCNEs), which are potential lncRNA precursors.

5. **Strengthen zebrafish as a model organism:** Improve lncRNA annotations in the zebrafish genome to expand its utility for studying lncRNA functions and evolutionary processes.

3. MATERIALS AND METHODS

3.1. Materials

3.1.1. Oligonucleotides for RNAseq library preparation and PCR reactions.

Table 1. Oligonucleotide sequences required for RNA-seq library preparation.

Blocker_PolyT_FOR	AAAAAAAAGCATCGCTGTCTCTTATACACATCTCCGAGCCCACGAGAC	Capture Protocol - CapTrap-seq Adapter blocker
Blocker_PolyT_REV	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGATGCTTTTTTTTTTTTTTTTVN	Capture Protocol - CapTrap-seq Adapter blocker
Blocker_UMI8_FOR	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNNNGTGGTATCAACGCAGAGTAC	Capture Protocol - CapTrap-seq Adapter blocker
Blocker_UMI8_REV	GTACTCTGCGTTGATACCACNNNNNNNNCTGTCTCTTATACACATCTGACGCTGCCGACGA	Capture Protocol - CapTrap-seq Adapter blocker
CapTrap_FOR	TCGTCGGCAGCGTC	CapTrap-seq library amplification Post-capture library amplification
CapTrap_REV	GTCTCGTGGGCTCGG	CapTrap-seq library amplification Post-capture library amplification
ISPCR oligo	AAGCAGTGGTATCAACGCAGAGT	TSO cDNA library amplification
oligodT16VN	P-TTTTTTTTTTTTTTTTVN	CapTrap-seq - Reverse Transcription step
Oligo-dT30VN	AAGCAGTGGTATCAACGCAGAGTACT$_{30}$VN	TSO - Reverse Transcription
Template Switching Oligo	AAGCAGTGGTATCAACGCAGAGTACATrGrG+G	TSO - Reverse Transcription
UMI16	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNNNNNNNNNNNGTGGTATCAACGCAGAGTA	CapTrap-seq - 2nd strand UMI primer

Table 2. Sequences of CapTrap-seq linkers

Name	Sequence (5'→3')
CapTrap-seq 5'-linker	GTGGTAUCAACGCAGAGUACGNNNNN-P P-CACCATAGTTGCGTCTCATG-P
CapTrap-seq 3'-linker	AAAAAGCAUCGCUGTCTCUTAUACACAUCUCCGAGCCCACGAGAC-P CGTAGCGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG

Table 3. PCR primers.

Name	Sequence (5'→3')	Application
eef1a1a_F1	GGTGGGATTGACAAAAGGACC	genomic DNA contamination check
eef1a1a_R1	CTTGTCCAGAACCCAGGCAT	genomic DNA contamination check

3.1.2. Kits and reagents

Table 4. Kits and reagents used in this work.

Kit/reagent	Company, Catalogue no.
Agencourt, RNAClean XP	Beckman Coulter, A63987
AMPure, XP Reagent	Beckman Coulter, A63881
Ligation Sequencing Kit	Oxford Nanopore Technologies, SQK-LSK109
myBaits Hybridization Capture for Targeted NGS Kit	Daicel Arbor Biosciences, myBaits Custom 80-100K, 300516.V5
ProNex® Size-Selective Purification System	Promega, NG2001
riboPOOL for Danio rerio	SiTools, 27DP-K012-000010
RNase-Free DNase Set	QIAGEN, 79254
SIRV-Set 4 (Iso Mix E0 /ERCC / Long SIRVs	Lexogen, 141.03
Template Switching RT Enzyme Mix	NEB, M0466
Vaccinia Capping System	NEB, M2080S

3.2. Methods

3.2.1. Maintenance of zebrafish lines and Ethical statement

Zebrafish wild type lines from AB background were maintained at the zebrafish core facility of the International Institute of Molecular and Cell Biology in Warsaw, Poland (License no. PL14656251), in compliance with institutional and national ethical guidelines for animal welfare. The fish were fed Gemma micro meal twice daily and maintained under a 14-hour light (8:00 to 22:00) and 10-hour dark (22:00 to 8:00) photoperiod.

3.2.2. Biological sample collection

Organs and Tissues from adult individuals

Adult AB wild-type individuals (older than 3 months) were used for organ and tissue dissection. Fish were first anesthetized with 120 mg/L of Tricaine methanesulfonate (MS-222) solution, then euthanized with an overdose of the same solution at 300 mg/L concentration (Sigma Aldrich, Catalog No. A5040-25G). Surface disinfection, rapid handling and RNase Free environment were ensured to protect dissected tissues and organs from degradation. Each collected biological sample (brain, heart, liver, ovary, or testis) was washed three times in sterile 1xPBS buffer (Invitrogen, Catalog No. AM9624) to remove potential contaminants such as blood, fat, or skin cells. Then, the excess PBS buffer was removed, and the biological samples were placed in RNase-free Eppendorf tubes containing 750 µl of TRIzol™ Reagent (Invitrogen, Catalog number: 15596026), followed by immediate homogenization with a Rotor-Stator Tissue Homogenizer (Omni International) for 20–30 seconds. Homogenized samples were incubated for 5 minutes at room temperature to allow complete dissociation of the nucleoprotein complexes, then frozen in liquid nitrogen and stored at -80°C until further RNA isolation.

Developmental stages

Embryos were obtained from the group spawning of male and female AB wild-type zebrafish. If specific early developmental stages (2-4 cells or shield stage) were needed, males and females were separated by a mesh, which was removed on the morning of the spawning day to allow the fish to breed. The obtained embryos were grown in an incubator at 28°C in egg water medium (1.5 ml sea salt stock added to 1 L distilled water, resulting in a final concentration of 60 µg/ml) until they reached

the desired developmental stage. The attainment of the intended developmental stage (2-4 cell stage, shield stage, or 28 hours post-fertilization) was monitored under the microscope. Zebrafish embryos were then washed three times in a sterile 1xPBS buffer to remove potential contaminants and debris. After removing the excess buffer, 50 embryos at the intended developmental stage were placed in RNase-free Eppendorf tubes containing 1 ml of TRIzol™ Reagent (Invitrogen, Catalog number: 15596026) and immediately homogenized with a Rotor-Stator Tissue Homogenizer (Omni International) for 20–30 seconds. To improve phase separation, the lysate was spun down for 5 minutes at 12,000 × g at 4°C, and the clear supernatant was transferred to a new tube. The samples were then incubated for 5 minutes at room temperature to allow complete dissociation of the nucleoprotein complexes, frozen in liquid nitrogen, and stored at -80°C until further RNA isolation.

3.2.3. RNA isolation

Biological samples stored in TRIzol™ Reagent at -80°C were thawed and equilibrated to room temperature. To each volume of TRIzol™ Reagent, 0.2 volumes of chloroform (Molecular Biology MP Biomedicals, Catalog No. ICN19400280) were added. The mixture was thoroughly shaken for 15 seconds and then incubated at room temperature for 5 minutes. The samples were then centrifuged at 12,000 × g for 15 minutes at 4°C. After transferring the aqueous phase to a new tube, 0.5 volumes of isopropanol (Merck, Catalog No. 1.09634.1000) were added per volume of TRIzol™ Reagent. For biological samples with an expected low RNA yield, 1 µl of RNase-free GlycoBlue™ Coprecipitant (15 mg/ml, Invitrogen, Catalog No. AM9515) was added to the aqueous phase as a carrier. The samples were thoroughly mixed and incubated at 4°C for 15 minutes, followed by centrifugation at 20,000 × g for 15 minutes at 4°C. After discarding the supernatant, the pellets were washed twice with 500 µl of 75% ethanol (Merck, Catalog No. 1.08543.0250) and spun down for 10 minutes at 20,000 × g at 4°C. Air-dried RNA pellets were then dissolved in an appropriate volume of RNase-free water. Any residual genomic DNA was removed by incubating the solution with RNase-Free DNase I (QIAGEN, Catalog No. 79254) at room temperature for 10 minutes, following the manufacturer's guidelines. The RNA samples were subsequently cleaned and concentrated using Agencourt RNAClean XP beads (Beckman Coulter, Catalog No. A63987) according to the 1.8× reaction volume protocol and eluted in UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™,

Catalog No. 10977015). The RNA concentration was measured using a Quantus Fluorometer (Promega), and purity was assessed with a Nanodrop Spectrophotometer (Thermo Fisher Scientific). RNA integrity was verified using the 2200 TapeStation system (Agilent Technologies). The RNA samples were stored at -80°C until further use.

3.2.4. RNA pooling and concentration

To increase the diversity of biological samples, RNA with the highest purity and integrity (RIN ≥ 8.5) was pooled and concentrated. Specifically, for developmental stages, RNA samples were combined from 1,000 zebrafish embryos per time point. For organ and tissue samples, RNA was pooled from 50 adult zebrafish, with 25 females and 25 males for brain, heart, and liver samples. Pooled RNA samples were thoroughly mixed with 0.1 volume of 3M RNase-free sodium acetate (pH 5.5, Invitrogen, Catalog No. AM9740) and 2.5 volumes of ice-cold 100% ethanol (Merck, Catalog No. 1.08543.0250). The samples were then precipitated overnight at -20°C, followed by centrifugation at 18,000 × g for 30 minutes at 4°C. The obtained pellets were washed twice with 500 µl of ice-cold 75% ethanol, with centrifugation at 18,000 × g for 10 minutes at 4°C each time. The ethanol was aspirated as much as possible, and the pellets were spun briefly (10 seconds at full speed) to remove any remaining ethanol. After air drying, the pellets were resuspended in nuclease-free water. RNA concentration and purity were measured using a Quantus Fluorometer (Promega) and a Nanodrop Spectrophotometer (Thermo Fisher Scientific), respectively. RNA integrity was assessed using the 2200 TapeStation system (Agilent Technologies). The RNA samples were aliquoted and stored at -80°C until further use.

3.2.5. Genomic DNA contamination check

Before proceeding, PCR validation was performed to check for genomic DNA contamination. The primers were designed using Primer-BLAST software to target a region of the zebrafish eef1a1a gene that contains an intron (GRCz11; chr13: 27,317,027-27,317,204; forward strand), which is spliced out in the processed RNA transcript. Consequently, amplicons obtained from the PCR reaction should differ in length: 93 bp for cDNA and 178 bp for genomic DNA. Then, cDNA was synthesized using RevertAid Reverse Transcriptase (200 U/µl) (Thermo Scientific™, Catalog No. EP0441). For each reaction, 500 ng of RNA, 0.5 µl of Oligo(dT) (100 µM), and 0.5 µl of Random Hexamer Primer (100 µM)

(Thermo Scientific™, Catalog No. SO142) were combined, with the final volume adjusted to 13 µl. To increase cDNA yield by reducing RNA secondary structure, the RNA and primer mixture was incubated at 65°C for 5 minutes, then immediately placed on ice. Next, the denatured sample was combined with RT Master Mix (4 µl of 5X Reaction Buffer, 2 µl of dNTP Mix (10 mM each), and 1 µl of RevertAid Reverse Transcriptase). The mixture was then incubated at 25°C for 10 minutes, at 42°C for 60 minutes, at 70°C for 10 minutes, and finally kept at 4°C.

The obtained cDNA sample was used for PCR amplification using Taq DNA Polymerase with ThermoPol® Buffer (NEB, Catalog No. M0267S). Specifically, 1 µl of undiluted cDNA was mixed with 2.5 µl of ThermoPol Reaction Buffer (10X), 0.5 µl of dNTP Mix (10 mM each), 0.5 µl of eef1a1a_F1 (10 µM), 0.5 µl of eef1a1a_R1 (10 µM), 0.125 µl of Taq DNA Polymerase, and the final volume was adjusted to 25 µl. cDNA amplification was performed with the following PCR cycling conditions: an initial denaturation at 95°C for 30 seconds, followed by 30 cycles of 30 seconds at 95°C, 60 seconds at 54°C, and 15 seconds at 68°C. A final extension was done at 68°C for 5 minutes, followed by a hold at 4°C. PCR products were visualized on a 3% UltraPure™ Agarose (Invitrogen™, Catalog No. 16500500) gel prepared in 1X TAE buffer and stained with SYBR™ Safe DNA Gel Stain (Invitrogen™, Catalog No. S33102). Electrophoresis was performed for 2.5 hours at 50V.

3.2.6. 5’-capping external spike-in controls

5’-capping of external spike-ins was performed following a previously described protocol (Carbonell-Sala et al. 2024). SIRV-Set 4 (Iso Mix E0 / ERCC / Long SIRVs, Lexogen, Catalog No. 141.03), an exogenous synthetic RNA control mix, underwent a 5’-capping step using the Vaccinia Capping System (NEB, Catalog No. M2080S). 15 µl of SIRV-Set 4 was incubated at 65°C for 5 minutes, then placed on ice for 5 minutes. The denatured RNA was combined with 2 µl of 10X Capping Buffer, 1 µl of GTP (10 mM), 1 µl of SAM (2 mM), and 1 µl of Vaccinia Capping Enzyme and incubated at 37°C for 2 hours. Finally, 5’-capped spike-ins were purified using Agencourt RNAClean XP beads (Beckman Coulter, Catalog No. A63987) following the 1.8X reaction volume protocol. After purification, the spike-ins were eluted with 15 µl of Nuclease-Free Water and incubated at 37°C for 5 minutes to enhance RNA recovery. The 5’-capped spike-ins were then stored at -80°C until further use.

3.2.7. cDNA sequencing library preparation

Four total RNA samples were used to benchmark the performance of different library preparation methods: total RNA from two adult tissues, heart and testis, and two total RNA samples from developmental stages, specifically the 2-4 cell stage and 28 hours post-fertilization (hpf). The RNA samples were quality controlled for concentration, integrity, and the absence of genomic DNA contamination.

The RNA was then processed using two different long-read library preparation methods, as detailed below:

3.2.7.1. CapTrap-seq library preparation

CapTrap-seq cDNA library preparation was performed following a previously described protocol (Carbonell-Sala et al., 2024; Carbonell-Sala et al., 2024) with the following changes: As input, 5 µg of RNA from each sample was mixed with SIRV-Set 4 (Iso Mix E0 / ERCC / Long SIRVs, Lexogen), which had been pre-5'-capped. For each reaction, 3 µl of a 1:100-diluted 5'-capped exogenous synthetic RNA control set was added. UMI16 oligonucleotide (100 µM stock, see Table 1) was used instead of UMI8 as the second-strand primer.

Additionally, second-strand cDNA samples were not subjected to a drying step but were directly used for Long and Accurate PCR (LA-PCR) with TaKaRa LA Taq (Cat. No. RR002M, Takara). To minimize PCR duplicates, the cDNA sample was split into two independent PCR reactions. Each reaction mix was prepared with 18.5 µl of nuclease-free water, 5 µl of 10x LA PCR Buffer II (Mg2+ plus), 8 µl of dNTPs mix (2.5 mM each), 2.5 µl of each primer (CapTrap_FOR and CapTrap_REV, in 10 uM stocks) (see Table 1), 0.5 µl of TaKaRa LA Taq (5 U/µl), and 13 µl of the second-strand synthesis product, resulting in a final volume of 50 µl per reaction.

The PCR cycling conditions were as follows: 30 s at 95°C for denaturation, 16 cycles of 15 s at 95°C, 15 s at 55°C, and 8 minutes at 68°C for amplification, followed by 10 minutes at 68°C and a hold at 4°C. The two PCR replicates were combined and purified with 1x AMPure XP beads, then resuspended in 21 µl of nuclease-free water. Samples were quantified with Qubit (Qubit 4 Fluorometer, Thermo Fisher Scientific) and quality checked with BioAnalyzer (Agilent 2100 Bioanalyzer, Agilent Technologies).

3.2.7.2. Library preparation using Template Switching approach

rRNA depletion

For library preparation, 5 µg of total RNA from each sample (Heart, Testis, 2-4 cell, 28 hpf) was mixed with 3 µl of 1:100-diluted, pre-5'-capped SIRV-Set 4 (Iso Mix E0 / ERCC / Long SIRVs, Lexogen). Next, each sample underwent rRNA depletion using the riboPOOL kit specifically designed for Danio rerio (SiTools, Cat. No. 27DP-K012-000010) according to the manufacturer's recommendations (Version riboPOOLKitManual_v6), with a few modifications. To protect RNA from degradation, 1 µl of RNasin Plus RNase Inhibitor (Promega, Catalog No. N2611) was added to each 13 µl of RNA sample. For optimal hybridization of oligonucleotides to rRNA molecules, the temperature was gradually decreased from 68°C to 37°C at a ramp rate of 0.1°C/s. After the ribodepletion step, RNA was cleaned up with Agencourt RNAClean XP beads (Beckman Coulter, Catalog No. A63987) using a 1.8X beads:sample ratio. The beads were washed three times with 70% ethanol, and the RNA was eluted with 22 µl of RNase-free water. Finally, 20 µl of the eluted sample was transferred to a clean tube. RNA samples were quantified with Qubit (Qubit 4 Fluorometer, Thermo Fisher Scientific) and. rRNA removal was verified using the 2200 TapeStation system (Agilent Technologies). The RNA samples were stored at -80°C until further use.

cDNA library preparation

4 µl of ribo-depleted RNA sample was used per reaction for first-strand synthesis using Template Switching RT Enzyme Mix (NEB, Cat. No. M0466), Template Switching Oligo (see Table 1) and anchored oligodT. Each sample was processed in duplicate, adhering closely to the manufacturer’s protocol. After completing first-strand synthesis, duplicates from each sample were pooled and purified using Agencourt RNAClean XP beads (Beckman Coulter, Catalog No. A63987) with a 1.8X beads-to-sample ratio. The beads were washed three times with 70% ethanol, and the cDNA was eluted with 41 µl of RNase-free water. Finally, 40 µl of the eluted sample was split into two aliquots.

A total of 40 µl was used for second-strand synthesis and amplification by Long and Accurate PCR (LA-PCR) with TaKaRa LA Taq (Cat. No. RR002M, Takara). To minimize PCR duplicates, the 40 µl was split into two independent PCR reactions. Each reaction mix was prepared with 14 µl of nuclease-free water, 5 µl of 10x LA PCR Buffer II (Mg2+ plus), 8 µl of dNTPs mix (2.5 mM each), 2.5 µl of ISPCR oligo (see Table 1) in 10 uM stock, 0.5 µl of TaKaRa LA Taq (5 U/µl), and 20 µl of the second-strand synthesis product, resulting in a final volume of 50 µl per reaction.

The PCR cycling conditions were as follows: 30 s at 95°C for denaturation, 9 cycles of 15 s at 95°C, 15 s at 65°C, and 8 minutes at 68°C for amplification, followed by 10 minutes at 68°C and a hold at 4°C. The two PCR replicates were combined and purified with 1x AMPure XP beads, then resuspended in 21 µl of nuclease-free water. Samples were quantified with Qubit (Qubit 4 Fluorometer, Thermo Fisher Scientific) and quality checked with BioAnalyzer (Agilent 2100 Bioanalyzer, Agilent Technologies).

3.2.8. cDNA size selection

The CapTrap-seq cDNA library was subjected to size-selection using the ProNex® Size-Selective Purification System (Promega, Catalog No. NG2001). The procedure was optimized to remove cDNA molecules shorter than 500 bp. Approximately 2 µg of the cDNA sample, adjusted to a final volume of 50 µl, was mixed with Promega beads at a 1:1.1 ratio and incubated at room temperature for 10 minutes on the HulaMixer™. Next, the beads with bound cDNA fragments longer than 500 bp were separated from the supernatant (SN) on a magnetic rack. The beads were then washed twice with freshly prepared 80% ethanol solution (without disturbing the beads), air-dried for 30 seconds, and resuspended in 20 µl of Elution Buffer. This mixture was incubated at 37°C for 10 minutes in a Thermoblock with gentle rotation. After incubation, the sample was placed on a magnet, and the solution containing the size-selected cDNA was recovered and transferred to a new Eppendorf tube.

Simultaneously, the supernatant containing cDNA fragments shorter than 500 bp was transferred to a new tube and purified by adding 2.0X volume of AMPure XP Reagent (Beckman Coulter, Catalog No. A63881), followed by incubation at room temperature for 10 minutes on a HulaMixer. The sample was then placed on a magnet, the supernatant was discarded, and the beads were washed twice with freshly prepared

80% ethanol solution (without disturbing the beads). Next, the beads were air-dried for 30 seconds and resuspended in 20 µl of Nuclease-Free water. The suspension was then incubated at 37°C for 10 minutes in a Thermoblock with gentle rotation. Immediately after incubation, the sample was placed on a magnet, and the solution containing the purified cDNA was recovered and transferred to a new Eppendorf tube. Quality check of the size-selection procedure was assessed using the Agilent 2100 Bioanalyzer system with the High Sensitivity DNA Kit (Agilent, Catalog No. 5067-4626). Additionally, the concentration and quantity of cDNA fractions obtained from the size-selection were measured using a Qubit fluorometer. The samples were stored at -20°C until further use. The cDNA fraction containing molecules longer than 500 bp (SS500bp) was used for sequencing with ONT platform.

3.2.9. Design of Zebrafish capture probes

The zebrafish capture probes were designed using the Ensembl version 104 annotation of the zebrafish genome (GRCz11). Target annotations were prepared in FASTA format and comprised the sets of features. When necessary, features were lifted over to the GRCz11 assembly. Features overlapping protein-coding gene loci were excluded. Intergenic lncRNAs were identified as genes with no transcript overlapping or located within 5 kb of any protein-coding gene. For novel regions predicted to encode lncRNAs, a 1 kb window centered around each region of interest was included. Predictions of zebrafish syntenic orthologues for human (gigaLNC) and mouse (NONCODE and GENCODE) lncRNAs were performed using the ConnectOR pipeline (https://github.com/Carlospq/ConnectOR).

Expression of candidate regions were quantified using publicly available RNA-Seq data from (Kaushik et al. 2013) and (Pauli et al., 2012) and top 30 highest-expressing regions were removed to prioritize rare transcripts. Moreover, ERCC spike-ins were included as external controls. Of the 92 ERCC sequences, the top 8 most concentrated were excluded and the remaining half (42) was selected to evenly span the concentration range.

In total, the design targeted 12,077 genomic regions, covering approximately 9.9 Mb of the zebrafish genome. All targets were compiled into a single FASTA file and submitted to Daicel Arbor Biosciences for probe design. The oligonucleotide probes were synthesized as a myBaits Custom 80-100K Library following the manufacturer's guidelines.

3.2.10. cDNA capture

LncRNA target enrichment was carried out using hybridization-based capture with the myBaits Hybridization Capture for Targeted NGS Kit and custom-designed RNA capture probes (Daicel Arbor Biosciences, myBaits Custom 80-100K, Item No. 300516.V5), following the manufacturer's protocol (Version 5.02) with following modifications.

For the procedure, 250 ng (in a maximum volume of 7 µl) of size-selected (SS500bp) CapTrap-seq cDNA libraries were used as input. To enhance target enrichment for cDNA molecules with inserts ranging from 1 to 10 kb, the Long Insert Protocol version was used. The Hybridization Reaction Setup was optimized for Zebrafish CapTrap-seq cDNA libraries. Specifically, the Most Taxa Capture Mix variant was used. Additionally, Block X reagent was substituted with adapter-blocking oligonucleotides customized for CapTrap-specific 5'- and 3'-linkers. The adapter-blockers were prepared as a mixture of UMI18 and PolyT blockers at a final concentration of 1 µg/µl. The sequences of the blocking oligos are detailed in Table 1. For target-bait hybridization and washing, a temperature of 62°C was used. All bead-washing steps were conducted in a water bath. Final post-capture library amplification was performed using KAPA HiFi HotStart ReadyMix (Roche, Catalog No. 7958927001, KK2601). Each PCR reaction utilized 10 µl of on-beads enriched libraries, combined with 10 µl of Nuclease-Free water, 25 µl of KAPA HiFi HotStart ReadyMix (2X), 2.5 µl of CapTrap_FOR primer (10 µM), and 2.5 µl of CapTrap_REV primer (10 µM). Two independent PCR reactions were assembled for each enriched library.

Post-capture library amplification was performed with the following PCR cycling conditions: denaturation for 3 minutes at 98°C, 18 cycles of 30 seconds at 95°C, 20 seconds at 60°C, and 10 minutes at 68°C for amplification, followed by a final hold at 4°C. Two PCR duplicates were pooled and placed on a magnet. The beads were discarded, and the supernatant containing the amplified cDNA molecules was transferred to a new tube.

The cDNA was purified by adding 1.0X volume of AMPure XP Reagent (Beckman Coulter, Catalog No. A63881), followed by a 15-minute incubation at room temperature on a HulaMixer. The samples were then placed on a magnet, the supernatant was discarded, and the beads were washed twice with freshly prepared 80% ethanol (without disturbing the beads). The beads were air-dried for 30 seconds and resuspended in 30 µl of Nuclease-Free water. The suspension was incubated at 37°C for 10 minutes in a Thermoblock with gentle rotation. Immediately after incubation, the sample was placed on a magnet, and the solution containing the purified cDNA was recovered and transferred to a new Eppendorf tube. The samples were stored at -20°C until further use for ONT sequencing.

3.2.11. ONT sequencing

Samples for Oxford Nanopore Technologies (ONT) MinION long-read sequencing were prepared from each TSO, CapTrap-seq and pre- or post-capture CapTrap-seq SS500bp library using the Nanopore Ligation Sequencing Kit based on 1D chemistry (SQK-LSK109), following the manufacturer's protocol (version ACDE_9064_v109_revG_23May2018). To achieve an optimal final loading concentration for MinION sequencing, 500 ng of each cDNA library was used as starting material. The cDNA molecules underwent end-repair and polyadenylation before Adapter Mix (AMX) ligation. The final ONT sequencing libraries consisted of double-stranded cDNA molecules with one adapter at each end. 150 fmols of each Nanopore sequencing library was loaded onto individual R.9.4.1 flow cells (FLO-MIN106D) and sequenced for at least 72 hours.

3.2.12. Data Analysis with the LyRic Pipeline

LyRic (https://github.com/guigolab/LyRic) is a versatile and automated transcriptome annotation and analysis workflow that facilitates the complete analysis of sequencing data, from mapping reads to the reference genome and building transcript models, to generating statistical files for a comprehensive quality check of the performed sequencing. It begins by mapping RNA sequencing data from ONT using Minimap2 to the zebrafish reference genome assembly danRer11, along with 176 Spike-In Controls (69 SIRVs, 92 ERCCs, and 15 long SIRVs). To compare LyRic output with reference annotations, a custom reference gene annotation file was created by integrating ENSEMBL gene annotations (v104), and Spike-In Controls annotations. Read aggregate

profiles along with annotated ENSEMBL genes were generated using the deeptools2 package. The compute Matrix and plotProfile functions were configured according to the LyRic documentation. The read-to-genome alignments were used to identify High Confidence Genome Mappings (HCGMs), which are characterized by four main criteria: (1) the presence of only canonical introns (for spliced reads), (2) the absence of suspicious introns potentially caused by RT template switching (for spliced reads), (3) a minimum average sequencing quality around the splice junctions (for spliced reads), and (4) the presence of a polyA tail in the read sequence (for unspliced reads). Finally, HCGM ONT reads were merged into a non-redundant set of transcript models (TMs) using *tmerge*. Next, LyRic transcript models (TMs) were merged with the ENSEMBL annotation (v104) using *tmerge*, identified novel loci with *buildLoci*, and these loci were classified as either intronic or intergenic. For full LyRic documentation, visit LyRic Documentation, and for detailed information, including the config file, refer to the LyRic data portal at LyRic GitHub.

3.2.13. RNAscope protocol

Whole-mount fluorescent in-situ hybridization using the RNAscope Multiplex Fluorescent v2 Assay kit (Advanced Cell Diagnostics) was performed according to the manufacturer's guidelines, with modifications as described by Gross-Thebing et al. (2014). Zebrafish embryos at 28 hours post-fertilization (hpf) in the AB background were manually dechorionated and then fixed overnight at 4°C in 4% paraformaldehyde (PFA) in 1xPBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7). The larvae were then washed for 5 minutes in 1xPBS and depigmented using a solution of 3% H_2O_2 and 1% KOH in distilled water (dH_2O) for 10 minutes at room temperature (RT), followed by a 5-minute wash in 0.1% PBT (0.1% Tween-20 in 1xPBS, pH 7.4). The embryos were gradually dehydrated through a series of increasing methanol (MeOH) concentrations (25%, 50%, 75%, 2× 100%) in 0.1% PBT (0.1% Tween-20 in PBS, pH 7.4), with 5-minute washes at each step. After the final MeOH wash, the embryos were stored at -20°C for at least one night before further use. The embryos were air-dried for 20-30 minutes at room temperature (RT) before proceeding with RNAscope-based signal amplification (Advanced Cell Diagnostics). Protease digestion of the embryos with Protease III was carried out for 20 minutes at room temperature (RT), followed by three 5-minute washes in 0.01% PBT (0.01% Tween-20 in PBS, pH 7.4). After removing the excess solution, target probe hybridization was carried out at 40°C

overnight to enhance probe penetration into deeper tissues. A DapB negative control probe and a custom-made probe (Dr-si-dkey-23i12.5) designed for *snhg1* (Advanced Cell Diagnostics) were used. Samples were treated with one probe at a time. After probe recovery, the embryos were washed three times for 15 minutes each in 0.2× SSCT (0.01% Tween-20, 3 mM NaCl, 0.3 mM TriNaCitratdihydrate, pH 7) at room temperature. An additional fixation step was performed using 4% paraformaldehyde for 10 minutes at room temperature. For RNA detection, all incubation steps were conducted in a 96-well plate using a HybEZ™ II Oven (ACD Biotechne) at 40°C. Hybridization with AMP1, AMP2, and AMP3 solutions, along with HRP signal development, was performed according to the protocol. Following each amplification (AMP) hybridization step, the embryos were washed three times for 15 minutes each with 0.2× SSCT. TSA VIVID 650 (Tocris Bioscience, Cat No. 7527/1 KIT) was used as the fluorophore at a 1:1,000 dilution. After HRP signal development and fluorophore incubation, the samples were washed three times for 10 minutes each with 0.2× SSCT at room temperature. Finally, the embryos were incubated overnight at room temperature with slow agitation in a ready-to-use DAPI solution (Advanced Cell Diagnostics). The larvae were mounted on glass slides with custom-made cavities using ProLong Gold Antifade Mountant (Thermo Fisher Scientific, Cat No, P10144). Images were captured using an LSM800 confocal laser scanning microscope (Zeiss).

4. RESULTS

The main goal of my PhD project was to improve lncRNA annotation in zebrafish. This involved refining existing models by integrating data from sources beyond Ensembl, identifying missing lncRNAs, and merging these resources to create the most comprehensive lncRNA repository. To achieve this, I **first** optimized the CapTrap-seq (Carbonell-Sala et al. 2024) full-length RNA sequencing method for zebrafish (Part A). **Then,** I combined it with the Capture Long-read Sequencing (CLS) approach developed by GENCODE (Lagarde et al., 2017) to establish a targeted full-length sequencing method, CapTrap-CLS (Kaur et al., 2024), which I used to generate high-throughput, manual-quality lncRNA annotations (Part B). **Lastly,** leveraging these high-quality models, I focused on the functional characterization of lncRNAs in zebrafish using advanced experimental techniques (Part C). In summary, my PhD project comprised three major components, each with distinct objectives and achievements, which are outlined in detail in the following sections.

PART A. Optimization of the CapTrap-seq full-length sequencing method for zebrafish

I. Benchmarking TSO and CapTrap-seq library preparation methods

First critical component of my PhD project was optimizing a full-length library preparation method for long-read RNA sequencing to allow for comprehensive analysis of the zebrafish transcriptome. The protocol needed to meet two essential criteria: it had to facilitate the investigation of full-length transcripts and effectively deplete rRNAs from the sequencing library. This latter criterion is especially crucial for less extensively studied model organisms such as zebrafish, where commercially available rRNA removal tools are often both ineffective and prohibitively expensive. To achieve this, I decided to evaluate two library preparation methods: a recently developed custom approach known as CapTrap-seq and a commercially available protocol based on Template Switching Oligo chemistry (NEB). CapTrap-seq is a method widely utilized by the GENCODE Consortium for generating comprehensive and accurate annotations of human and mouse genomes. This method has proven particularly effective in enriching sequencing libraries with full-length molecules, thereby addressing the issue of 5'-end incompleteness (Carbonell-Sala et al. 2024). Therefore, it appeared to be an ideal method for optimization in zebrafish.

RNA input samples

The benchmarking of these two library preparation methods was conducted using a range of zebrafish samples. To obtain a comprehensive evaluation of method performance, I tested TSO and CapTrap-seq approaches on total RNA samples derived from two adult organs (heart and testis) and two developmental stages (2-4 Cell and 28 hours post-fertilization) (Figure 4.1).

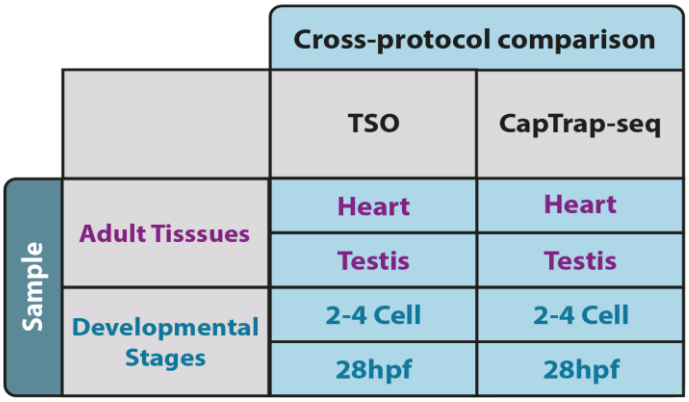


Figure 4.1 Biological samples used for full-length library preparation method comparison. *Two adult tissues, heart and testis, as well as two developmental stages, 2-4 cell and 28 hpf, were utilized for cross-protocol comparisons to evaluate the quality of CapTrap-seq. The vertical blue line represents the cross-protocol comparisons between the two sequencing library preparation methods: TSO and CapTrap-seq.*

Both protocols aim to generate full-length cDNA libraries; therefore, high RNA integrity is crucial. Consequently, only samples with RNA Integrity Number (RIN) values of ≥8.5 and high purity were used for RNAseq library preparation (Figure S1).

Moreover, background RNA contamination with genomic DNA (gDNA) is often overlooked in RNA-seq studies. Such contamination can lead to false results or transcript model artifacts, including an increased proportion of monoexonic reads. To ensure the reliability of the obtained annotations, it is crucial to use gDNA-free RNA samples. To achieve this, I included a DNase digestion step during RNA isolation protocol. Additionally, the efficiency of gDNA removal was assessed by performing a PCR reaction to amplify a fragment of the zebrafish *eef1a1a* gene containing an intron (GRCz11; chr13: 27,317,027-27,317,204; +). This fragment is spliced out in the processed RNA transcript but remains in the case of gDNA (Figure S2). The tests confirmed that the proposed method was efficient in detecting residual gDNA amounts (Figure S3) and that all RNA samples were free of genomic DNA contamination (Figure S4) and suitable for sequencing library preparation.

cDNA library preparation and sequencing

Sequencing libraries using the TSO and CapTrap-seq approaches were synthesized from the same RNA aliquot to minimize variability associated with different RNA batches. Moreover, for both protocols, 5 µg of total RNA was used as the starting material, which was spiked with 5'-capped external RNA control mixes prior to the procedure. The primary difference between the CapTrap-seq and TSO protocols is that the latter does not include a step specifically designed to enrich 5'-capped transcripts. Additionally, standard library preparation using cDNA synthesis with an oligo(dT) primer alone is insufficient for the effective removal of rRNA molecules, leading to a significant fraction of reads mapping to this RNA biotype. Therefore, I incorporated an rRNA depletion step using a zebrafish-specific kit prior to initiating the library preparation protocol to ensure the effective removal of rRNA transcripts.

The cDNA profiles of the RNA-seq libraries exhibited distinct sample-specific patterns and demonstrated high reproducibility across library replicates. Notably, clear differences were observed between the cDNA profiles produced by the two library preparation methods (Figure S5), indicating variations in how these approaches target the zebrafish transcriptome. Next, all prepared cDNA libraries were sequenced using the Oxford Nanopore Technologies platform, with each sample run on a separate ONT flow cell. Significant inter-run variability was observed between samples and tested protocols in terms of sequencing read yield (Figure 4.2). Similar observations of variability across ONT runs have already been reported, suggesting that this diversity may arise from differences in library composition or from the quality and number of available pores on the ONT flow cell (Ni et al., 2023).

For the analysis of the obtained sequencing data, the Lyric pipeline (Carbonell-Sala et al., 2023; https://github.com/guigolab/LyRic) was used. This tool, developed for the GENCODE Consortium by Julien Lagarde, is designed to enable detailed and comprehensive analysis of RNA-seq data, encompassing tasks from read mapping to the reference genome, assessment of transcript model completeness, to annotation construction (Figure 4.3). Additionally, it generates detailed statistical files that facilitate a comprehensive comparison between library preparation protocols or biological samples. LyRic was designed primarily for use with human or mouse samples, but it has no limitations for use with other organisms, including zebrafish.

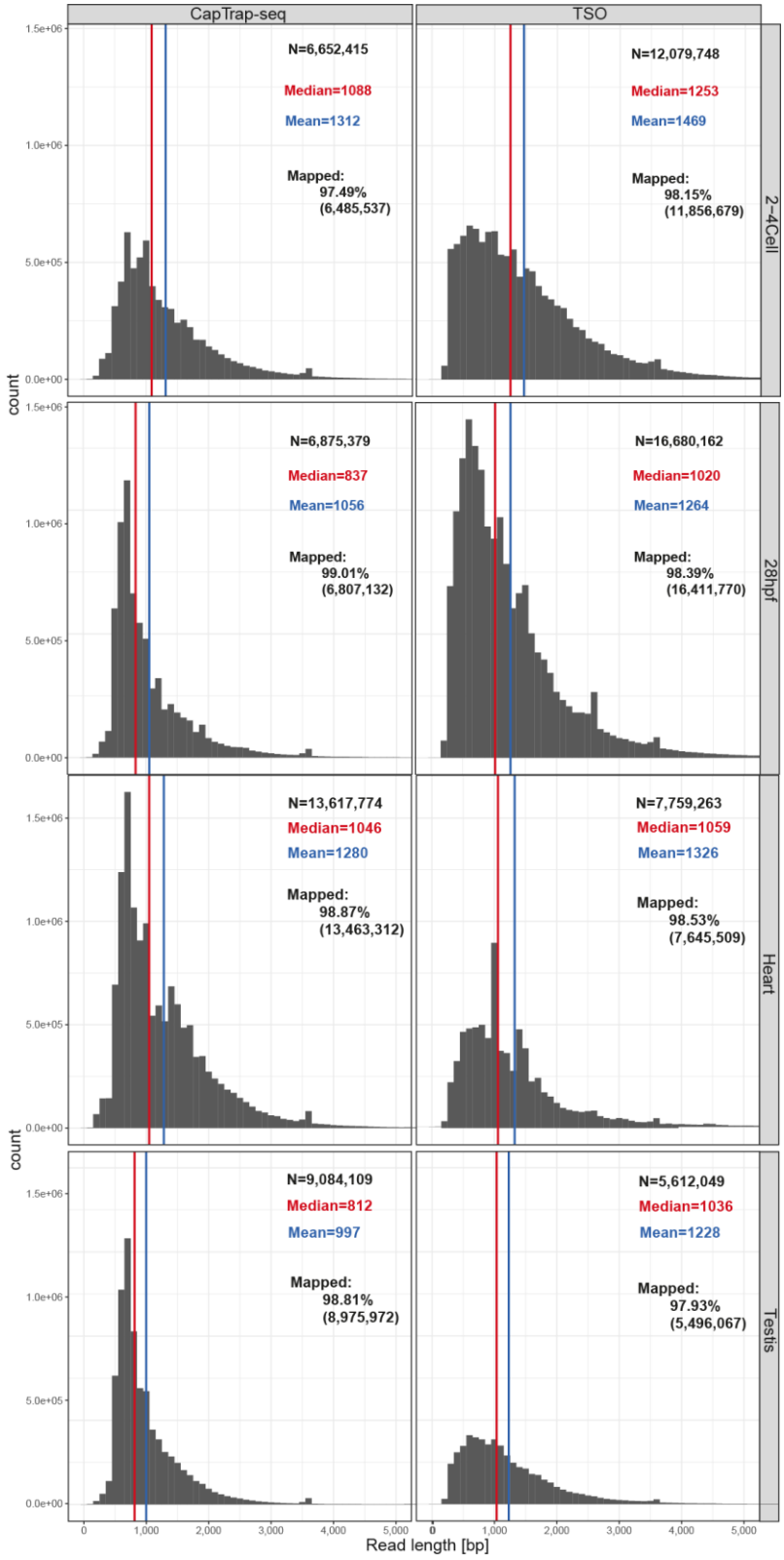


Figure 4.2. The length distribution of raw long-read ONT reads for each protocol across biological samples. *The total number of reads (N), along with the median and mean read lengths (represented by red and blue vertical lines, respectively), and mapping rate are displayed in the top right corner.*

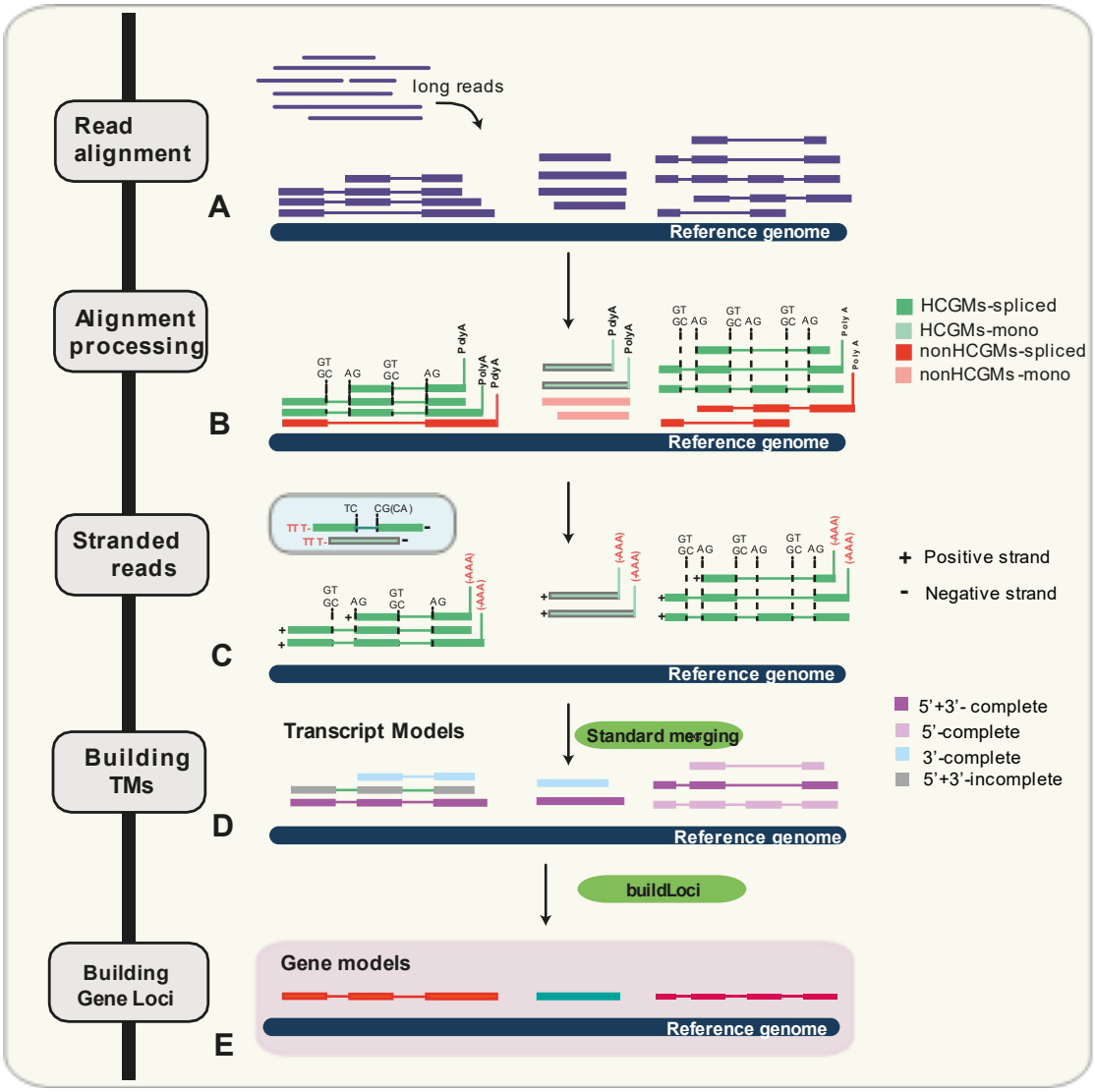


Figure 4.3. The framework of the LyRic pipeline. *LyRic incorporates five main steps: (A) Aligning the reads and mapping them to the reference genome; (B) Processing alignments to select only high-quality reads; (C) Assessing the DNA strand from which cDNA was generated; (D) Constructing transcript models and evaluating their 5' and 3' completeness; (E) Constructing gene models.*

ONT reads quality assessment

ONT sequencing reads were subjected to thorough quality assessment using the LyRic pipeline. The results demonstrated consistently high mapping rates across all samples, with values ranging from 97% to 99% (Figure 4.2). However, reads generated by the TSO method were, on average, 200 base pairs longer than those from CapTrap-seq. This variation in read length distribution was already apparent at the cDNA profiling stage (Figure S5), where the difference between TSO and CapTrap-seq reads was even more pronounced, with lengths ranging between the protocols from 600 to even 800 nucleotides.

Sequencing errors can significantly affect transcript annotation by introducing artificial bases, which may result in inaccurate identification of splice junctions and isoform structures. As a part of this step, I assessed how the various library preparation protocols impacted the number of errors generated during Nanopore sequencing. The analysis demonstrated that, while both methods produced a similar number of deletions and insertions, the TSO protocol generated a higher number of mismatches compared to CapTrap-seq. These results are consistent with previous reports indicating that template switching methodologies are prone to generating sequencing errors, such as mismatches (Zhang et al., 2022; Verwilt et al., 2023). This discrepancy may further impact data analysis and compromise the accuracy of transcript model reconstruction (Figure 4.4).

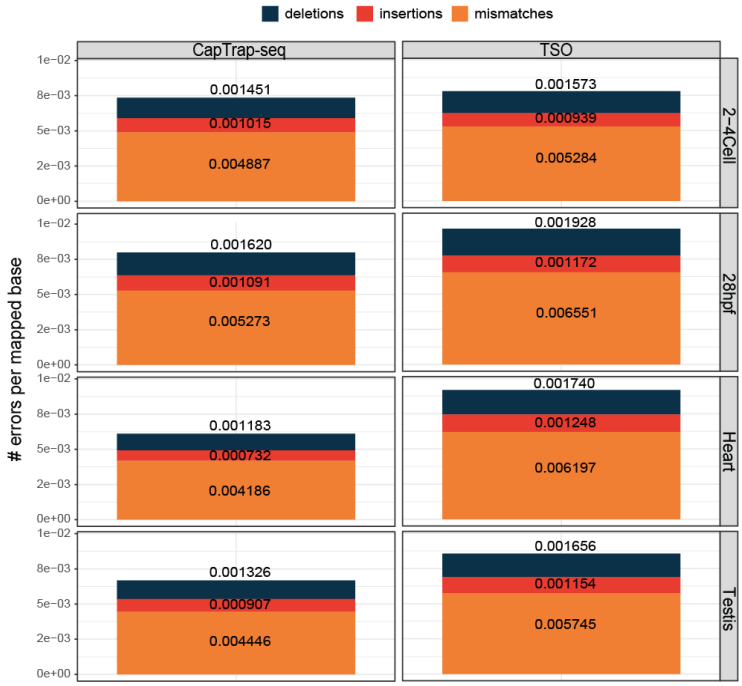


Figure 4.4. The error rate across CapTrap-seq and TSO sequencing library preparation protocols for zebrafish samples. *Different colors denote the types of errors: deletions (black), insertions (red), and mismatches (orange).*

Given the observed differences in read length distribution, I investigated how the CapTrap-seq and TSO protocols influenced the detection of external spike-in molecules. The external RNA control mix added to the samples included SIRVs (Spike-In RNA Variants) with lengths ranging up to 3 kb. I categorized the detection rates of these spike-ins into three groups: fully detected (end-to-end), partially detected (partial), or not detected (absent). The analysis revealed that the efficiency of end-to-end SIRV detection was slightly lower for CapTrap-seq samples, particularly for the longest SIRVs exceeding 2 kb (Figure 4.5).

This reduced detection efficiency for longer SIRVs can likely be attributed to the CapTrap-seq protocol, which, being a multistep process, may favor the enrichment of shorter cDNA molecules, potentially leading to the loss of longer ones during processing. In contrast, the TSO protocol demonstrated a higher proportion of partially detected SIRVs compared to CapTrap-seq (Figure 4.5). This aligns with the TSO protocol's design, which excludes a 5'-cap selection step, enabling the detection of fragmented cDNA molecules. It is important to note that an error occurred during the preparation of the heart CapTrap-seq sample, where uncapped SIRV spike-ins were mistakenly added. This error negatively affected their detection, resulting in the detection of only a limited number of spike-ins below 1 kb in this sample, compared to the corresponding sample prepared using the TSO protocol (Figure 4.5).

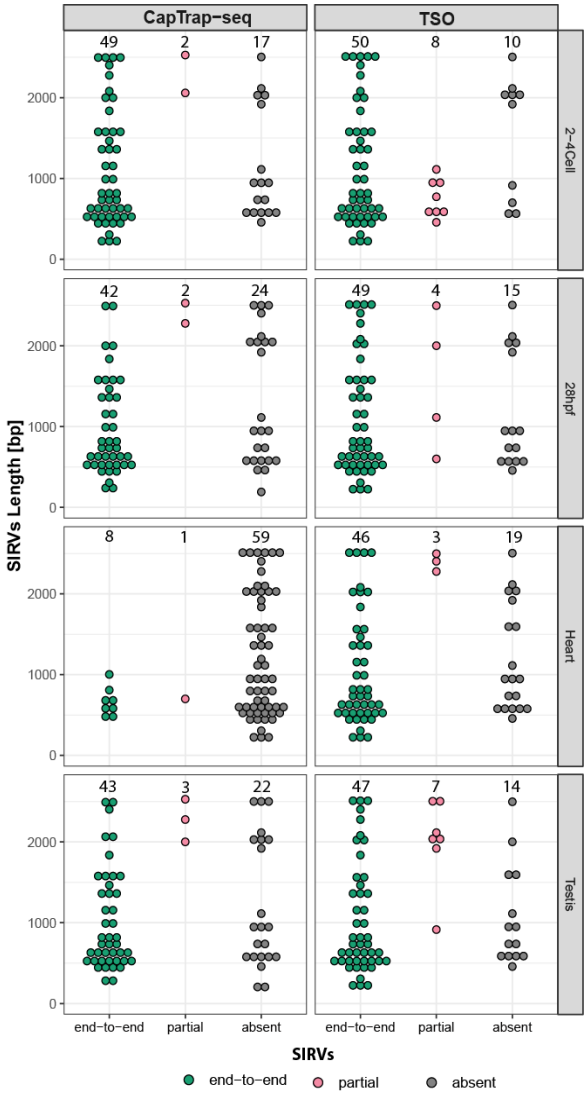


Figure 4.5. Detection of SIRVs based on their length. *Three detection levels were identified: end-to-end (green), partial (red), and not detected/absent (gray). The black numbers at the top represent the total count of SIRVs for each detection level.*

Further differences between the protocols were observed in the detection of various RNA biotypes after mapping the reads to ENSEMBL-annotated genes. Both CapTrap-seq and TSO protocols, by incorporation of polyA selection step, targeted a polyadenylated fraction of the transcriptome. Consequently, it is unsurprising that the majority of reads mapped to protein-coding genes. Moreover, both protocols similarly targeted lncRNAs. However, the TSO approach generated a comparatively larger fraction of non-exonic reads, particularly in the testis sample (Figure 4.6).

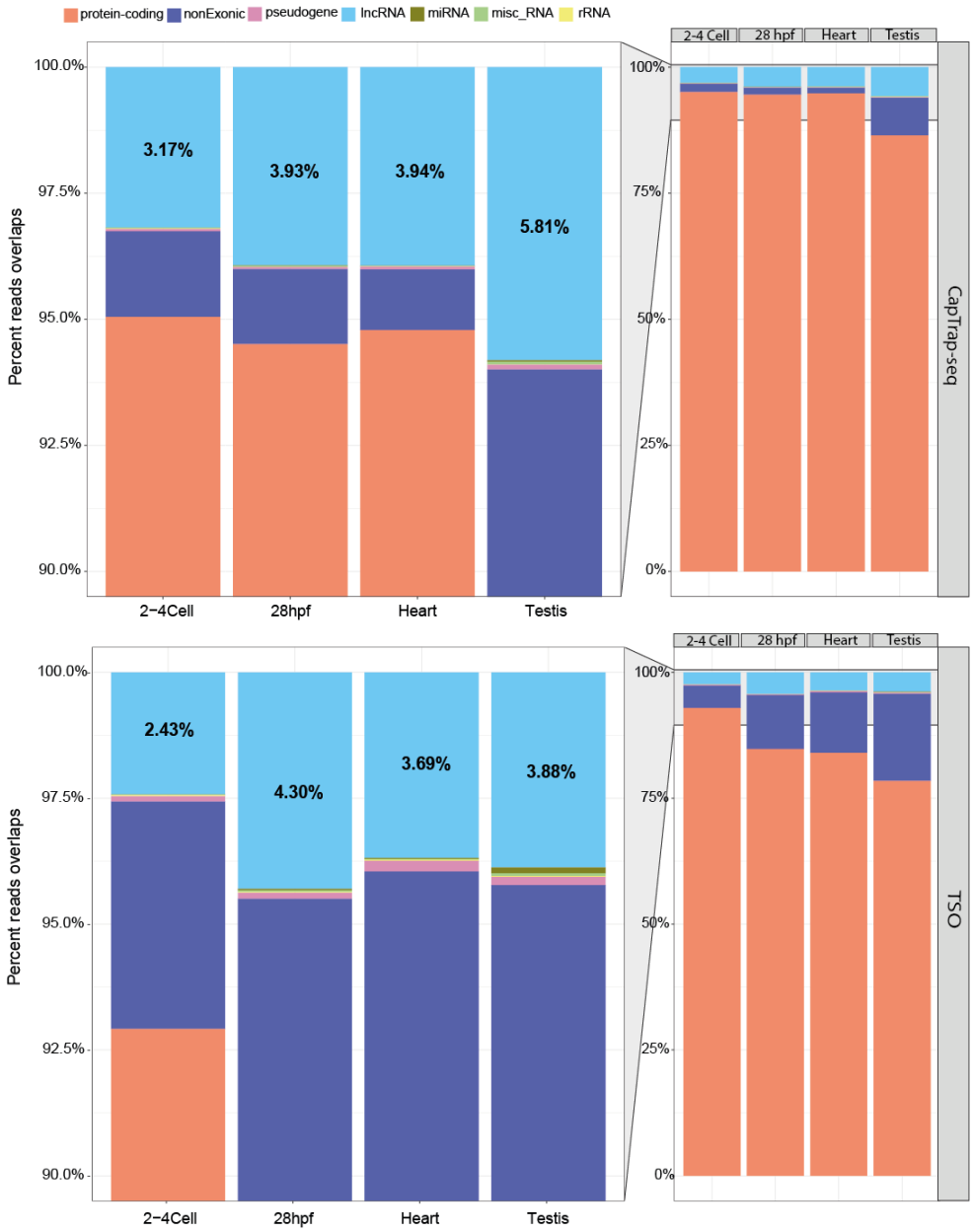


Figure 4.6. GENCODE gene biotype detection by ONT sequencing *of A) CapTrap-seq and B) TSO protocols. The stacked bar plots show the percentage of raw reads mapping to various annotated ENSEMBL gene biotypes. Different colors represent the ENSEMBL gene biotypes: protein-coding (orange), ribosomal RNAs (yellow), pseudogenes (pink), miscellaneous RNA (green), miRNAs (olive green), non-exonic (dark blue), and lncRNAs (light blue).*

Upon closer examination of the fraction of reads mapping to the rRNA biotype, it was observed that for both protocols, this proportion was significantly below 1% (Figure 4.7). This is a noteworthy outcome, especially considering that in most zebrafish transcriptomic studies, the rRNA fraction often exceeds 5%. Interestingly, despite the inclusion of an organism-specific rRNA removal step in the TSO protocol, the fraction of rRNA-originating reads was substantially higher, exceeding that of the CapTrap-seq method by more than tenfold (Figure 4.7).

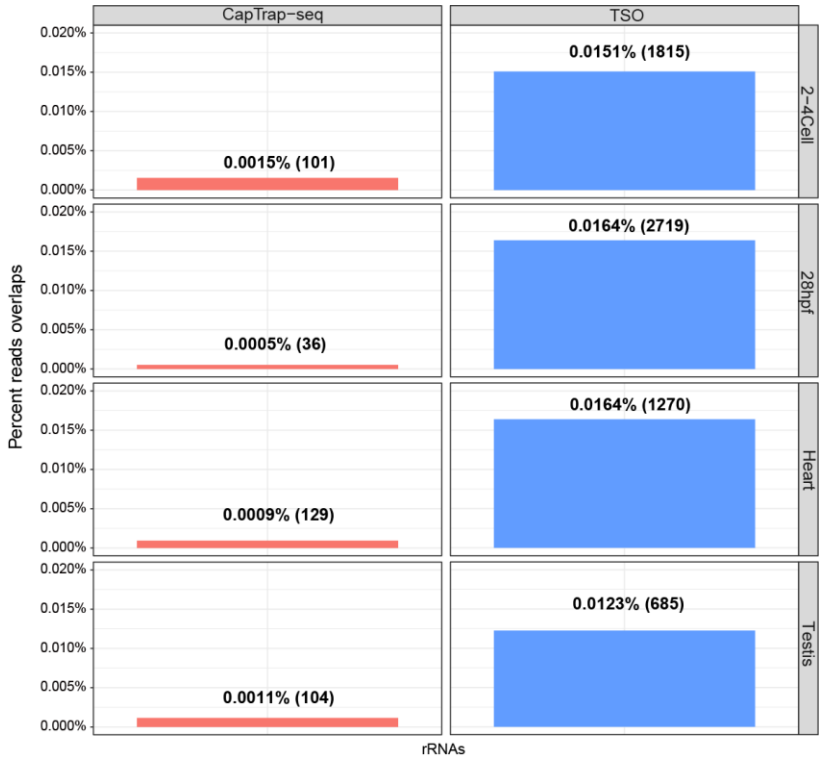


Figure 4.7. Comparison of rRNA biotype detection across zebrafish samples and library preparation methods. *The proportion of raw reads mapping to rRNA, along with the absolute raw read numbers in parentheses, is indicated above each bar plot.*

The next quality control step involved assessing the fraction of polyadenylated reads. Both the CapTrap-seq and TSO methods employ polyadenylated transcript selection using oligo(dT) reverse transcription primers; therefore, both protocols are expected to enrich sequencing libraries with a high fraction of polyadenylated reads. However, it was observed that CapTrap-seq produced a lower fraction of polyadenylated reads compared to the TSO method, with the exception of the heart sample (Figure 4.8).

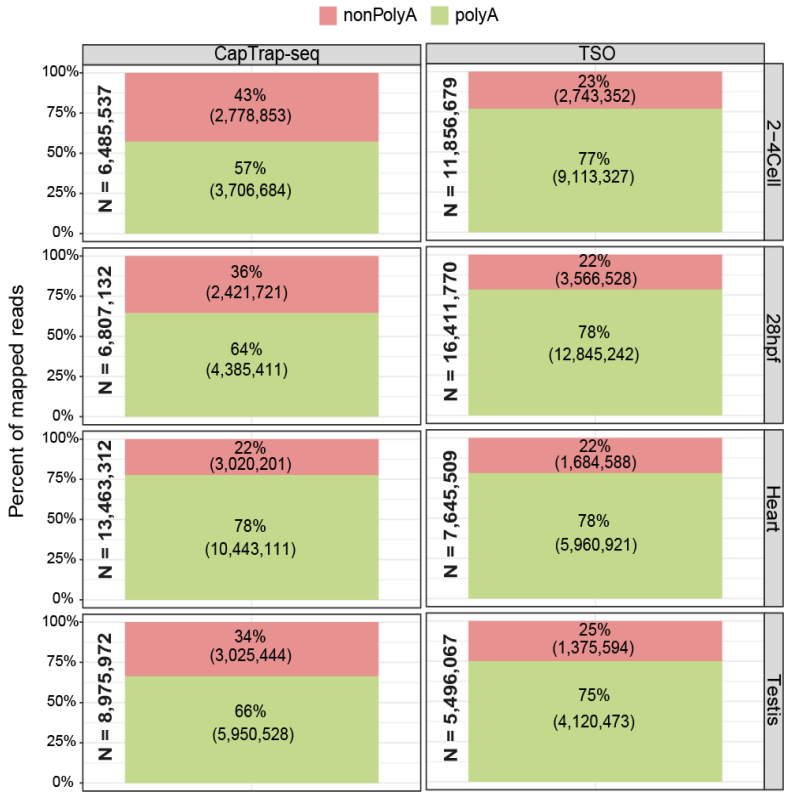


Figure 4.8. Detection of polyadenylated reads across zebrafish samples and library preparation methods. *Proportion of poly(A) (green) and non-poly(A) (red) ONT reads across zebrafish samples for CapTrap-seq and TSO library preparation methods.*

Next, the mapped ONT reads were utilized to call High Confidence Genome Mappings (HCGMs) using the LyRic pipeline. The approach for calling HCGMs differs between spliced and unspliced reads. HCGMs for spliced reads consist solely of those with canonical and high-quality splice junction sequences, while HCGMs-unspliced reads require the presence of a detectable polyA tail (Figure 4.3 B).

Reads generated with the TSO protocol were of slightly lower quality compared to CapTrap-seq. CapTrap-seq produced a greater fraction of High Confidence Genome Mappings (HCGMs), with approximately 60% of these being spliced, compared to only around 40% for TSO (Figure 4.9). In contrast, TSO generated a higher proportion of unspliced reads, including both HCGMs and non-HCGMs. It is widely accepted that spliced reads are more reliable than unspliced ones, as the latter can often result from transcriptional artifacts or fragmented transcripts therefore CapTrap-seq appears as better approach in this context.

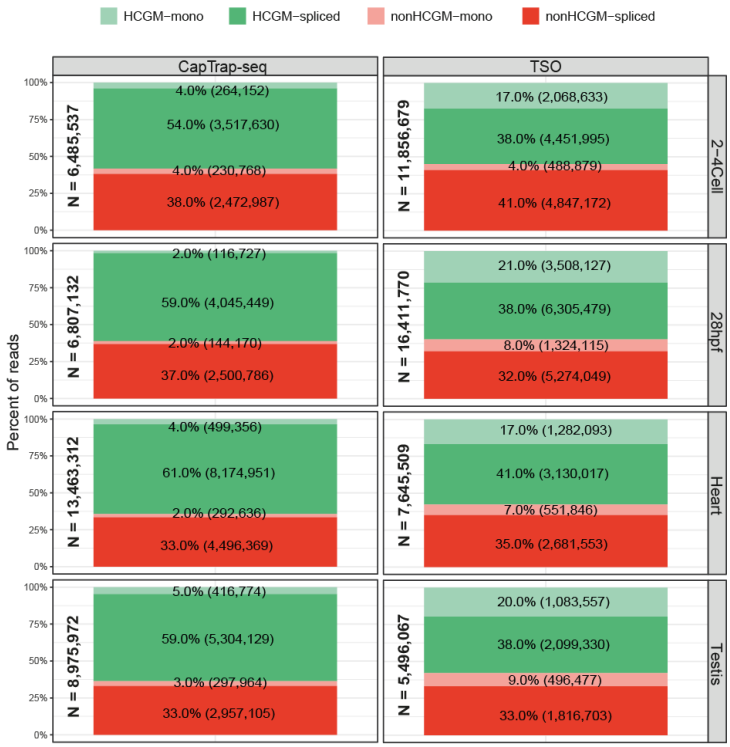


Figure 4.9. The proportion of detected High-Confidence Genome Mappings (HCGMs). *Four main classes include: unspliced HCGMs (light green), spliced HCGMs (dark green), unspliced non-HCGMs (light red), spliced non-HCGMs (dark red).*

Transcript model reconstruction and assessment of 5’ and 3’-end completeness

Next, HCGM-spliced and unspliced reads were utilized to construct transcript models using the Lyric pipeline, which were subsequently evaluated for 5' and 3'-end completeness (Figure 4.3 D). Lyric employs distinct methodologies to assess the evidence supporting predicted transcript ends. The completeness of the 3'-end was determined by the presence of an unmapped polyA tail, while 5'-end completeness was evaluated based on the proximity to Cap Analysis of Gene Expression (CAGE) tags from the DANIO-CODE dataset.

The annotation created from CapTrap-seq data consisted of over 80% complete transcript models (TMs) supported by both 5' and 3'-ends, compared to less than 70% for the TSO protocol (Figure 4.10). It also demonstrated a higher percentage of spliced transcript models (TMs), with approximately 85% for CapTrap-seq compared to around 60% for TSO. Furthermore, 70% of the spliced TMs were supported by both 5' and 3'-ends in CapTrap-seq, in contrast to roughly 45% for TSO. Notably, TSO produced a greater fraction of monoexonic transcripts, many of which were 5'-incomplete (Figure 4.10). Overall, CapTrap-seq significantly outperformed the TSO method in recalling full-length transcript models.

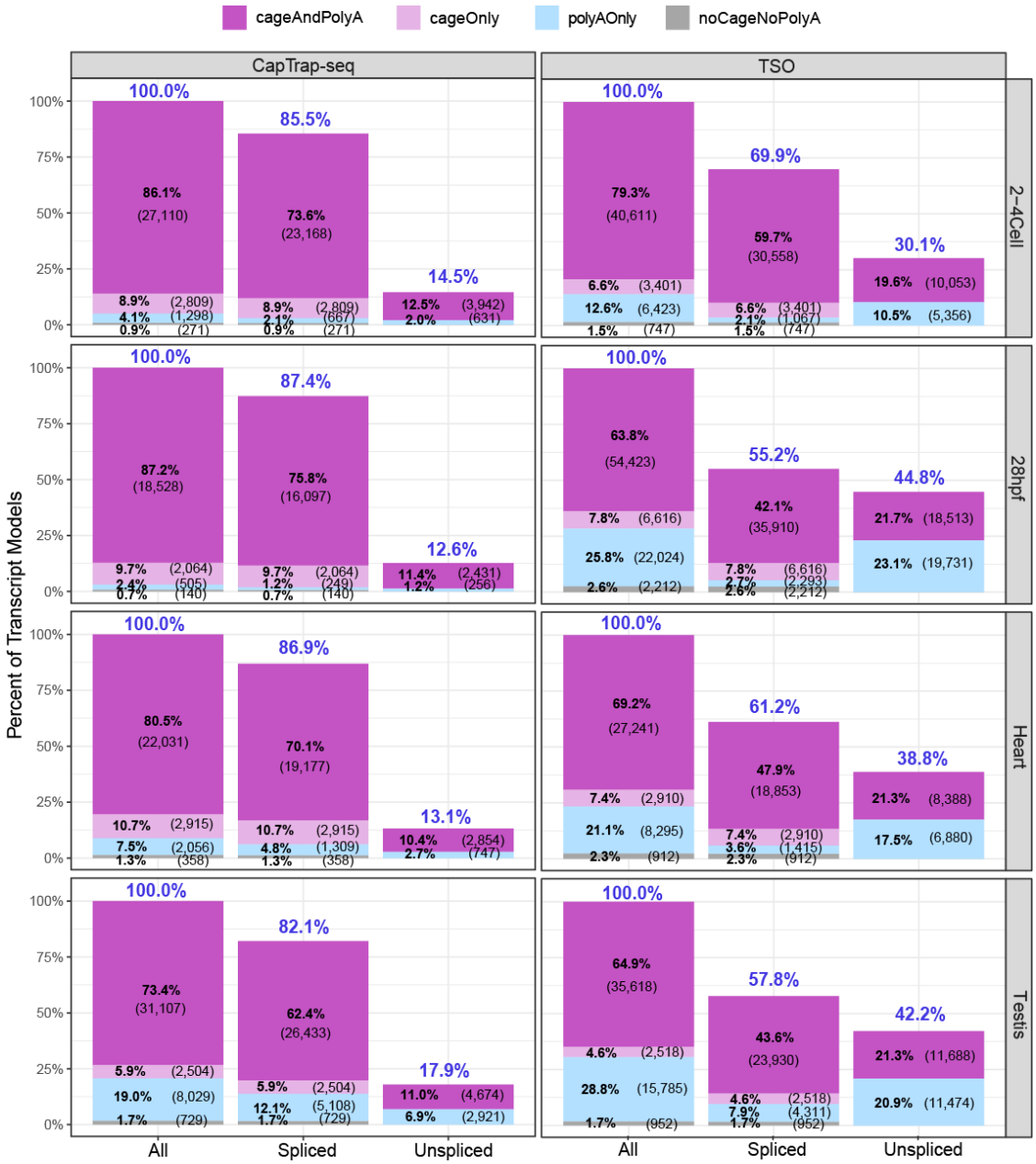


Figure 4.10. Detection of full-length transcript models (FL-TMs) *among all, spliced and unspliced TMs. Four categories of transcript model (TM) completeness are represented: Grey for incomplete TMs, Sky blue for 3'-complete TMs, Light pink for 5'-complete TMs, and Purple for full-length TMs.*

The difference in length distribution between the CapTrap-seq and TSO protocols remained evident at the transcript model (TM) level. The median TM length obtained by CapTrap-seq was approximately 100 bp to 350 bp shorter compared to TSO for both All TMs and spliced TMs, as well as for All and full-length TMs (Figure 4.11). However, the analysis shows that the TSO protocol generates a larger fraction of unsupported transcript models compared to CapTrap-seq, which is particularly evident in the All TMs group (Figure 4.11).

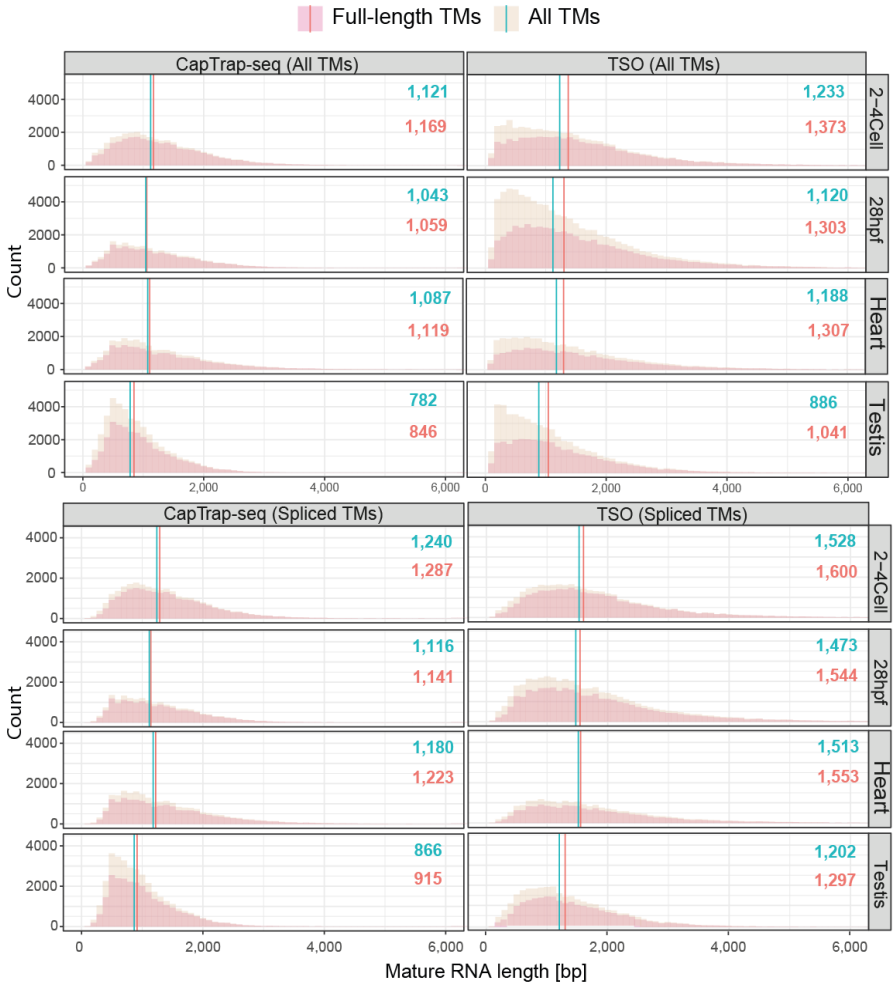


Figure 4.11. Transcript Model length distribution. *Length distribution of complete (peach) and All (beige) transcript models, encompassing both All and spliced models is depicted. The median read length for All (turquoise) and FL (red) TMs is indicated in the top right corner.*

II. Application of cDNA size selection to enhance the performance of CapTrap-seq

cDNA library preparation and sequencing

CapTrap-seq has proven to be an effective method for the high-quality reconstruction of the zebrafish transcriptome. However, its primary limitation lies in its preference for shorter molecules. To address this issue, I aimed to incorporate a size-selection step into the CapTrap-seq protocol to enhance its performance. Various size-selection methods are available; however, for my PhD project, I focus on a bead-based approach. A size-selection cutoff of 500 base pairs was established to strike a balance between increasing the representation of longer cDNA molecules in the sequencing library and maintaining sufficient sensitivity to annotate shorter transcripts (Figure 4.12).

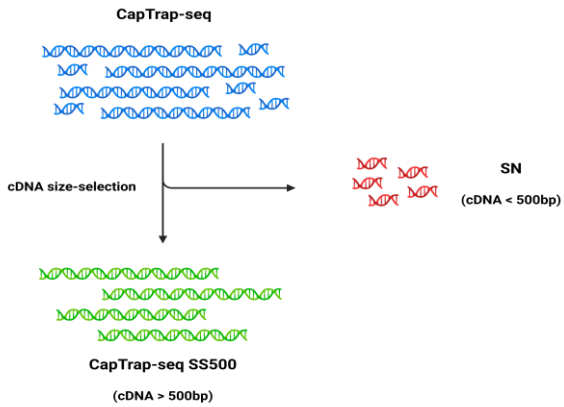


Figure 4.12. Schematic representation of the size-selection step in the CapTrap-seq protocol *(blue). This step involves the removal of cDNA fragments shorter than 500 bp (indicated as SN, red), resulting in a size-selected sequencing library called CapTrap-seq SS500 (green).*

Consequently, I applied the size-selection step on comparable samples used in the CapTrap-seq versus TSO benchmarking, specifically heart, testis, 2-4 Cell stage, and 28 hpf. The application of SS500 exhibited sample-specific effects, likely related to the differential proportion of 500 bp molecules in the initial libraries tested. While size-selection improved the average cDNA length for heart samples (increasing by approximately 150 bp) and 28 hpf samples (increasing by about 200 bp), the effect on cDNA length for the testis and 2-4 Cell samples appeared to be negligible (Figure 4.13).

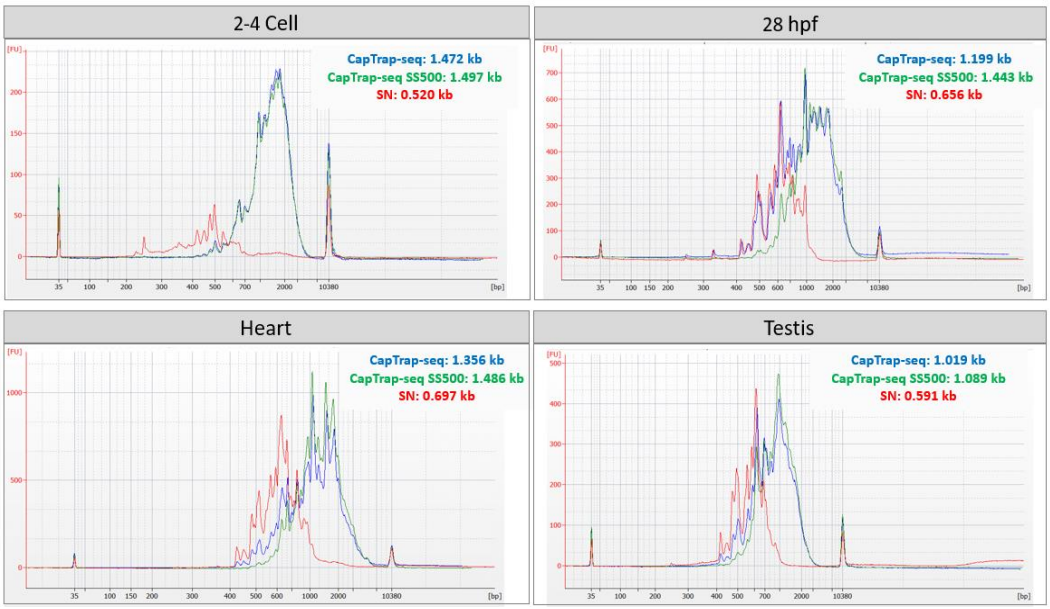


Figure 4.13. cDNA profiles across zebrafish samples before and after size selection step. *CapTrap-seq library before size selection (blue), after size selection (green) and the fraction removed (red).*

At the read level, the size-selection step improved the median read length only for the 28 hpf and testis samples, while it resulted in a minimal decrease for the 2-4 cell and heart samples (Figure 4.14).

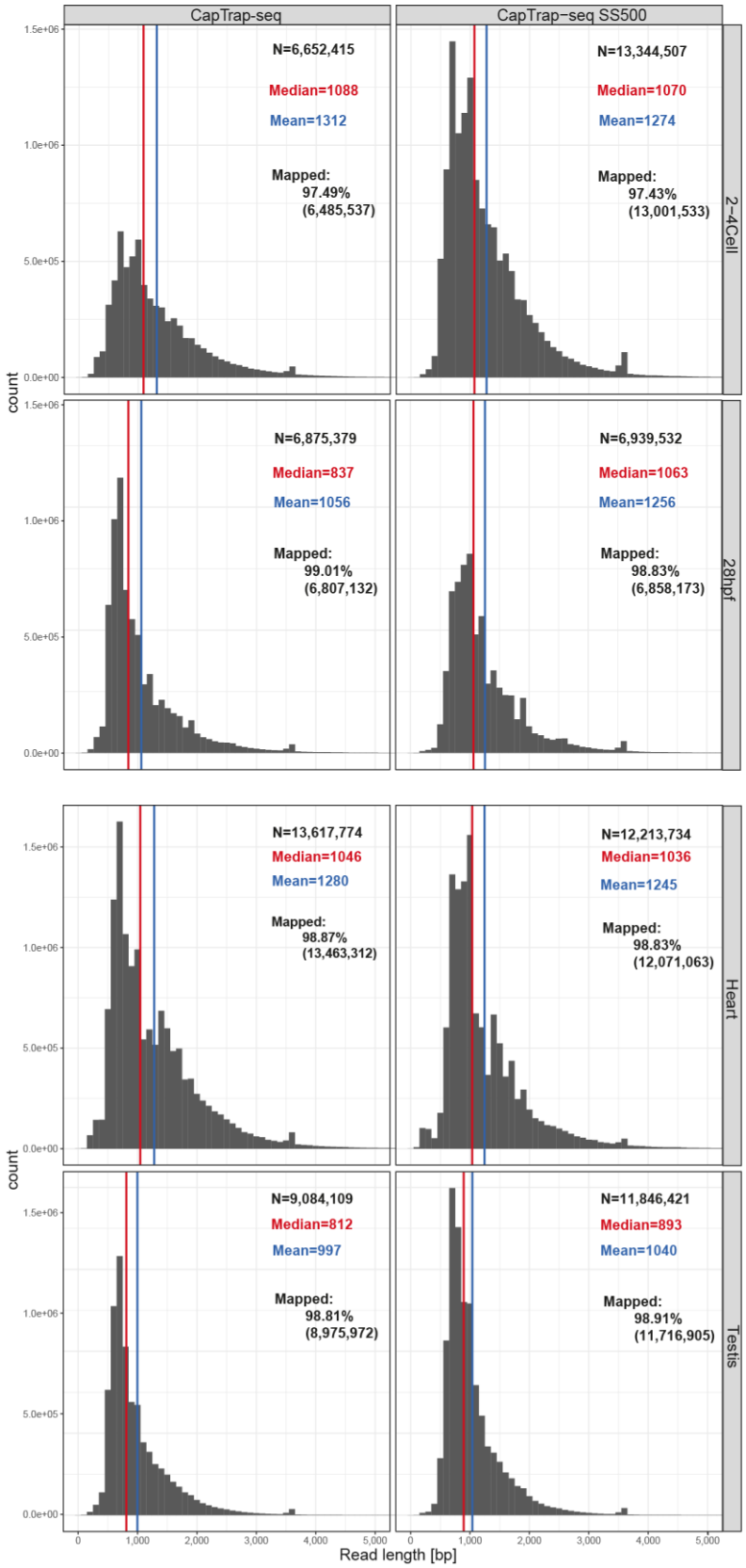


Figure 4.14. The length distribution of raw long-read ONT reads *for the standard CapTrap-seq method and the size-selection (SS500) modality, across biological samples, is depicted. The total number of reads (N), as well as the median and mean read lengths (indicated by red and blue vertical lines, respectively), along with the mapping rate, are displayed in the top right corner.*

ONT reads quality assessment

Next, I investigated the effect of size-selection on detection of externally provided SIRVs. As expected, size selection with a 500 bp cutoff negatively impacted the detection of SIRVs shorter than 500 bp (Figure 4.15). However, I observed a slight improvement in the detection of SIRVs in higher length ranges, particularly those above 2 kb, for CapTrap-seq SS500bp samples (Figure 4.15). As expected, the inclusion of uncapped SIRVs in the heart CapTrap-seq sample also impacted their detection in the corresponding SS500 sample, with the majority remaining undetected (Figure 4.15).

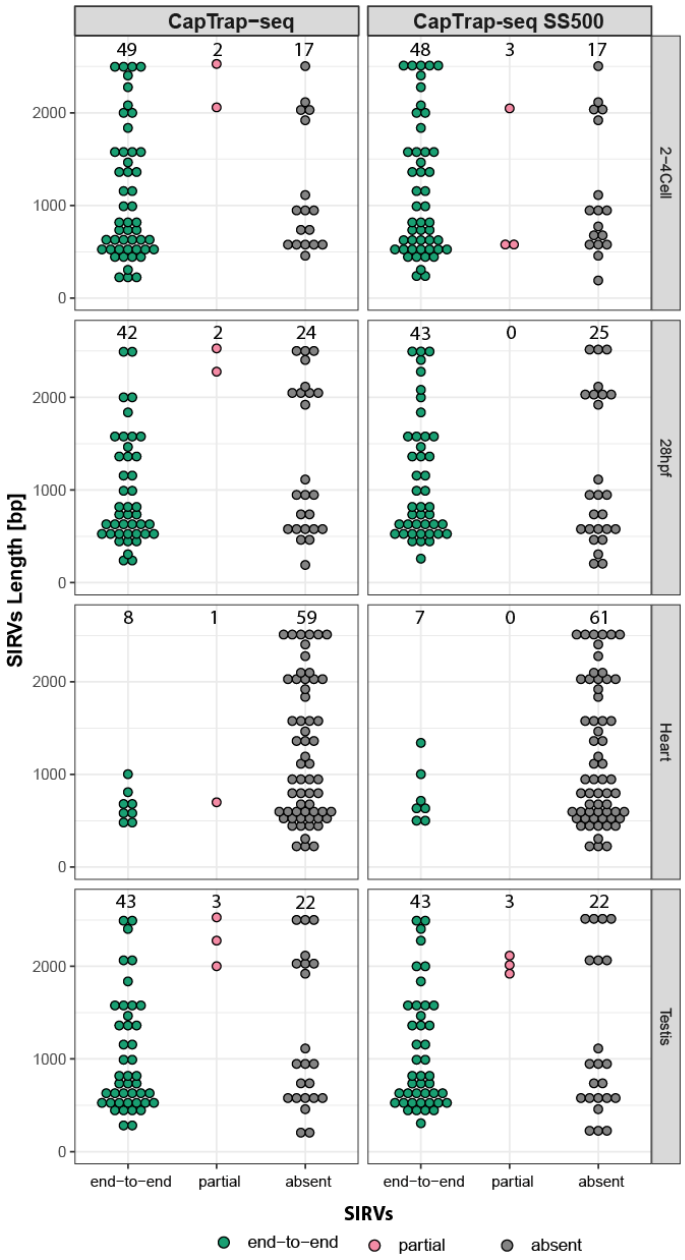


Figure 4.15. Detection of SIRVs based on length. *Three detection levels were identified: end-to-end (green), partial (red), and not detected/absent (gray). The black numbers at the top represent the total count of SIRVs for each detection level.*

Another limitation of CapTrap-seq compared to TSO is its lower proportion of polyadenylated reads; therefore, I aimed to investigate the effect of size-selection on the detection of reads with polyA tails. Notably, size-selection with a 500 bp cutoff resulted in a substantial increase in the fraction of polyadenylated reads (Figure 4.16), nearly approaching the values achieved with the TSO approach.

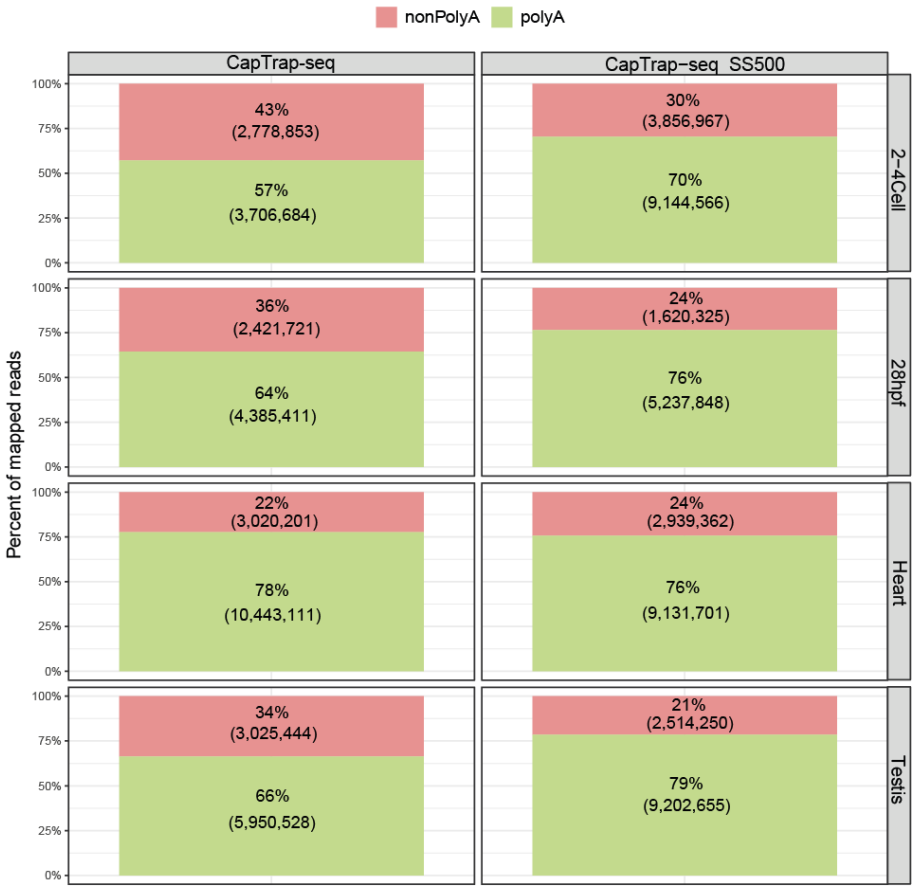


Figure 4.16. Detection of polyadenylated reads across zebrafish samples and library preparation methods. *The proportion of poly(A) (green) and non-poly(A) (red) ONT reads across zebrafish samples is shown for both the standard CapTrap-seq method and the SS500 modality.*

Transcript model reconstruction and assessment of 5’ and 3’ end completeness

Notably, size-selection had no negative effect on either the fraction of 5' and 3'-complete transcript models or the percentage of spliced TMs (Figure 4.17), thereby maintaining the high-quality performance of CapTrap-seq in enriching full-length spliced transcript models.

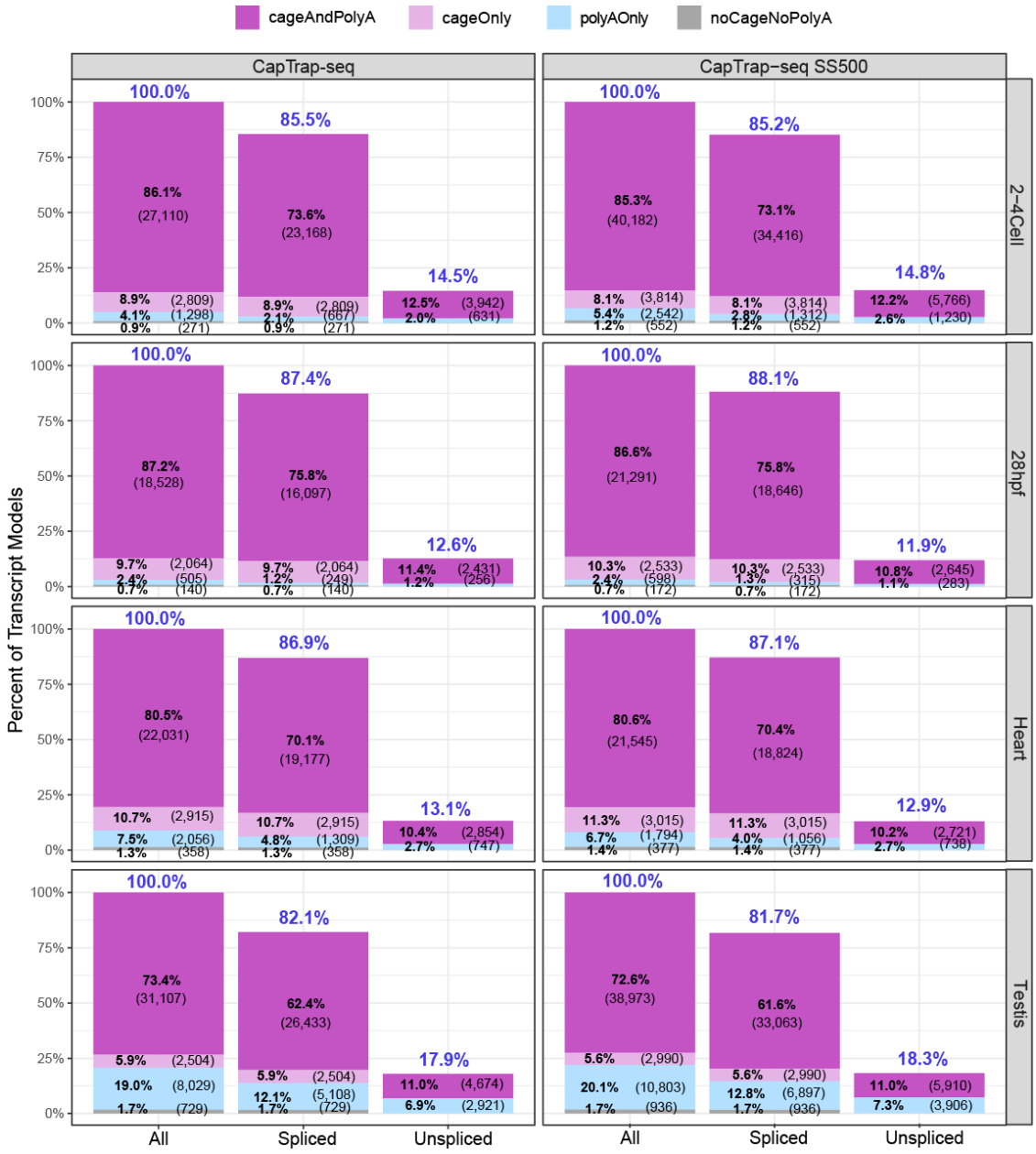


Figure 4.17. Detection of full-length transcript models (FL-TMs) *among all, spliced and unspliced TMs. Four categories of transcript model (TM) completeness are represented: Grey for incomplete TMs, Sky blue for 3'-complete TMs, Light pink for 5'-complete TMs, and Purple for full-length TMs.*

Finally, I aim to investigate the effect of size-selection on the median length of generated transcript models. With the exception of the 2-4 cell sample, size selection with a 500 bp cutoff resulted in an increase in the median length of transcript models generated by CapTrap-seq. This improvement in length was observed for both all transcript models and for the spliced ones (Figure 4.18). Therefore, the application of size-selection helped to mitigate limitations of the CapTrap-seq protocol, ultimately improving the sequencing of longer cDNA molecules.

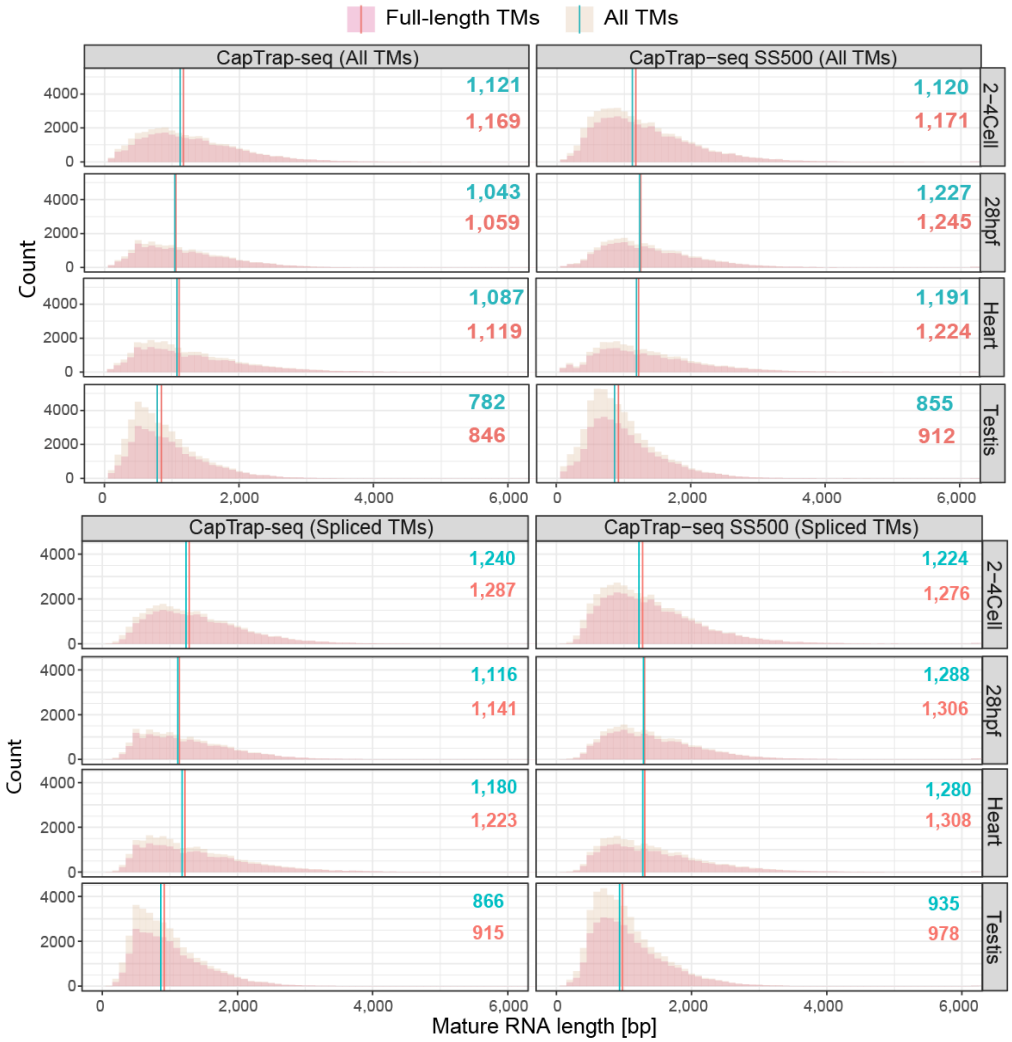


Figure 4.18. Transcript Model length distribution. *Length distribution of complete (peach) and All (beige) transcript models, encompassing both All and spliced models. The median read length for All (turquoise) and FL (red) TMs is indicated in the top right corner.*

The Effect of Library Preparation Method on ENSEMBL Annotation Extension

I next examined how different library preparation methods influenced the extension of the zebrafish ENSEMBL annotation by comparing 5'+3'-complete annotations generated with CapTrap-seq, CapTrap-seq SS500 or TSO to Ensembl reference. At first glance, the TSO method appeared more effective in identifying novel loci within the intergenic space (Figure 4.19, all TMs); however, this was likely attributable to a higher proportion of unspliced transcript models located in unannotated genomic regions. When focusing exclusively on spliced transcript models, no significant differences were observed in the fraction of intergenic transcript models between the TSO and CapTrap-seq protocols, with the exception of the testis sample. In this case, the fraction was approximately twofold higher in CapTrap-seq and CapTrap-SS500 samples compared to TSO (Figure 4.19, spliced TMs).

Figure 4.19 Effect of used library preparation method on ENSEMBL annotation extension. *The generated annotations were compared to the reference (ENSEMBL), with transcript models classified into seven categories: Included, Equal, Overlaps, Antisense, Intronic, Extends, and Intergenic. A deeper purple color corresponds to a greater degree of novelty of the transcript model.*

A more detailed analysis revealed that, although the application of the TSO method facilitated the identification of a higher number of novel loci, only 20–40% of these loci were classified as spliced. Additionally, the majority of these loci were deemed incomplete (Figure 4.20). In contrast, the implementation of CapTrap-seq, particularly with the use of the size-selection modality (S500), enabled the identification of a greater number of spliced and full-length loci compared to the TSO method (Figure 4.20).

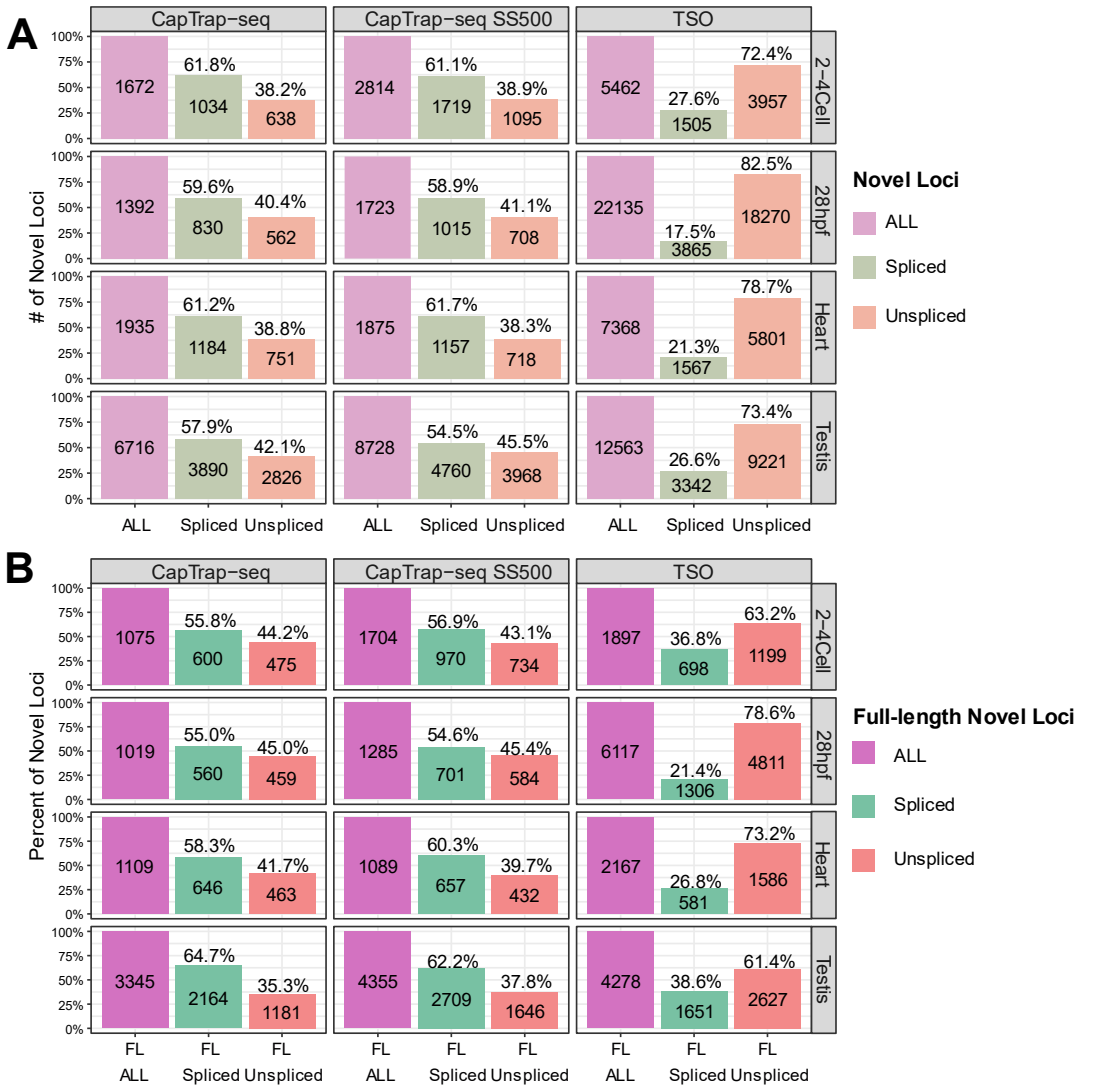


Figure 4.20. Novel loci identification. *Effect of the library preparation method on (A) ALL and (B) full-length (FL) novel loci detection. The novel loci identified, shown in purple, were further classified into spliced (represented in green) and unspliced (shown in orange).*

Several instances highlighted the efficiency of CapTrap-seq in extending annotations by identifying full-length transcripts. This capability is particularly evident in biologically significant protein-coding genes such as *lhx9*, a LIM homeobox transcription factor essential for neuron fate specification and thalamus development (Peukert et al., 2011) and *tbx5b*, a T-box transcription factor crucial for zebrafish heart development and pectoral fin formation (Pi-Roig et al., 2014). For both genes, CapTrap-seq identified novel TSSs that were absent in both the Ensembl annotation and TSO samples (Figure 4.21 A and B). Furthermore, these novel TSSs and TTSs were supported by CAGE tags and 3P-seq peaks from DANIO-CODE, further validating their accuracy. The novel transcription start site (TSS) identified in the CapTrap-seq sample for the *lhx9* gene was also supported by DANIO-CODE consensus promoters and detected in other CapTrap-seq samples, including the brain sample (Figure 4.21 A).

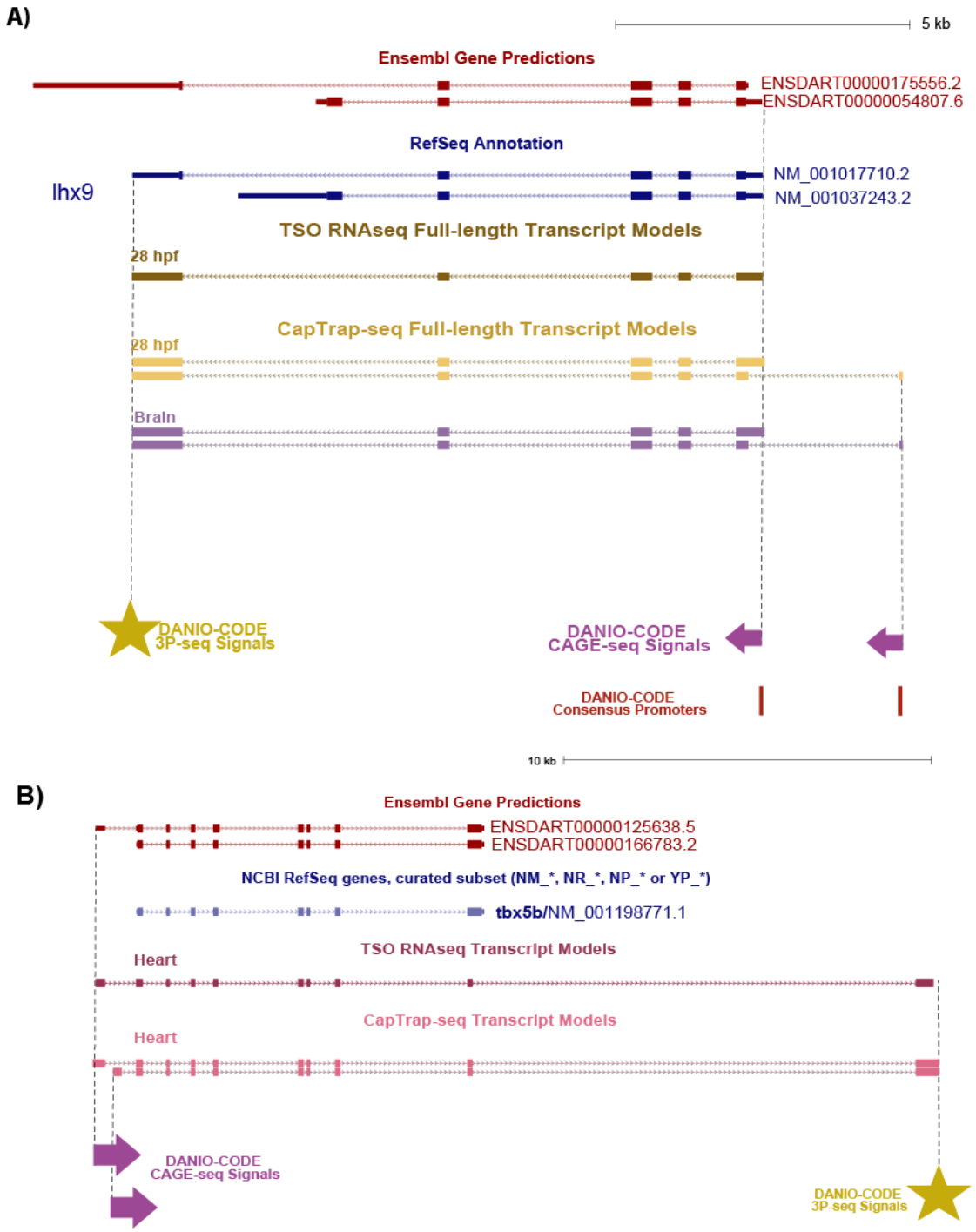


Figure 4.21. Novel transcript models identified for A) *lhx9* and B) *tbx5b* protein-coding genes. *5'-end support from CAGE-seq is indicated by a purple arrow, while 3'-end support from 3P-seq data is represented by a yellow star.*

Improvements in 5'-end completeness were also noted for lncRNAs. CapTrap-seq more accurately reconstructed the transcription start site for *mir9-3hg* compared to Ensembl or TSO approach. The TSS identified by CapTrap-seq aligns well with CAGE tags and consensus promoter regions from DANIO-CODE, further validating its accuracy (Figure 4.22 A).

Additionally, while both CapTrap-seq and TSO successfully reconstructed the transcript model for *mir1-1hg* and identified novel transcript isoform for this gene, the TSO method produced a significant proportion of unspliced transcript models, the majority of which were incomplete at either the 5' or 3' ends (Figure 4.22 B).

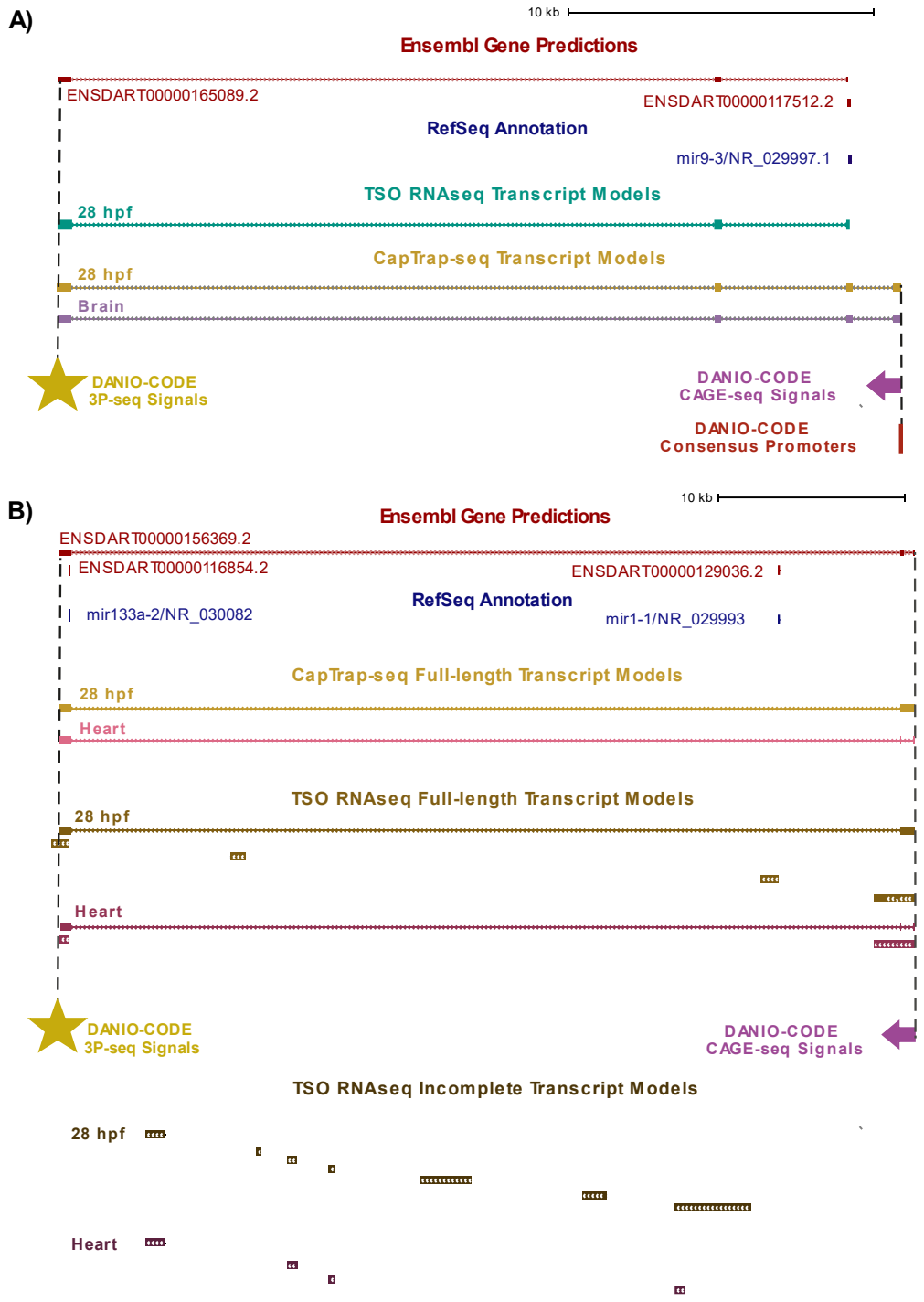


Figure 4.22. Updated annotation of A) *mir9-3hg* and B) *mir1-1hg* lncRNAs. *5'-end support from CAGE-seq is indicated by a purple arrow; 3'-end support from 3P-seq is represented by a yellow star.*

PART B. Annotation of lncRNAs using the CapTrap-CLS approach

Please note: Positionally conserved genes were identified using the in-house developed tool ConnectOR, designed by Prof. Barbara Uszczyńska-Ratajczak and Prof. Rory Johnson. The code was initially developed by Dr. Carlos Pulido (Johnson's lab) and later modified by Dr. Sasti Gopal Das (Uszczyńska-Ratajczak's lab).

I. ConnectOR – a synteny-based method for predicting lncRNA orthologs

Currently, we still lack a solid foundation to confidently determine which lncRNAs are functional. Since functional characterization of lncRNAs requires investigating their loss of function and the resulting phenotypic changes individually, a more efficient method for prioritizing potentially functional lncRNAs for validation is needed. Decades of research on protein-coding genes have shown that evolutionary conservation is a reliable indicator for identifying biological functions, offering a promising approach for lncRNA studies as well. However, due to their modular architecture, lncRNAs may not exhibit homology at the primary sequence level. Analyzing structural similarity is also challenging, given the vast number of lncRNAs in mammalian genomes, incomplete annotations, and the presence of fragmented molecules.

One of the most intriguing, yet underexplored, aspects of lncRNA biology is positional conservation, where the genomic context and relative transcriptional orientation are preserved across species. To facilitate the detection of potentially functional lncRNAs in a sequence-independent manner across human, mouse, and zebrafish genomes, we collaborated with Rory Johnson's lab to develop ConnectOR (Connect Orthologous RNAs), a synteny-based approach for predicting orthologous lncRNAs. Instead of comparing individual lncRNA sequences, ConnectOR is designed to detect conserved sequence blocks (Figure 4.23 A) in a species-agnostic manner through two-way comparisons. ConnectOR uses standard genome annotations as input and operates in two modes: exonic or genic overlap (Figure 4.23 B). To evaluate the performance of orthologous gene prediction, both modes were tested using GENCODE annotations for human and mouse genomes as an input. The exonic overlap mode proved to be more effective for identifying orthologous genes compared to the genic overlap mode (Figure 4.23 C), as lifting longer genic sequences across species is more challenging than with shorter exonic sequences. Therefore, exonic overlap mode was chosen for detecting positionally conserved genes across different organisms.

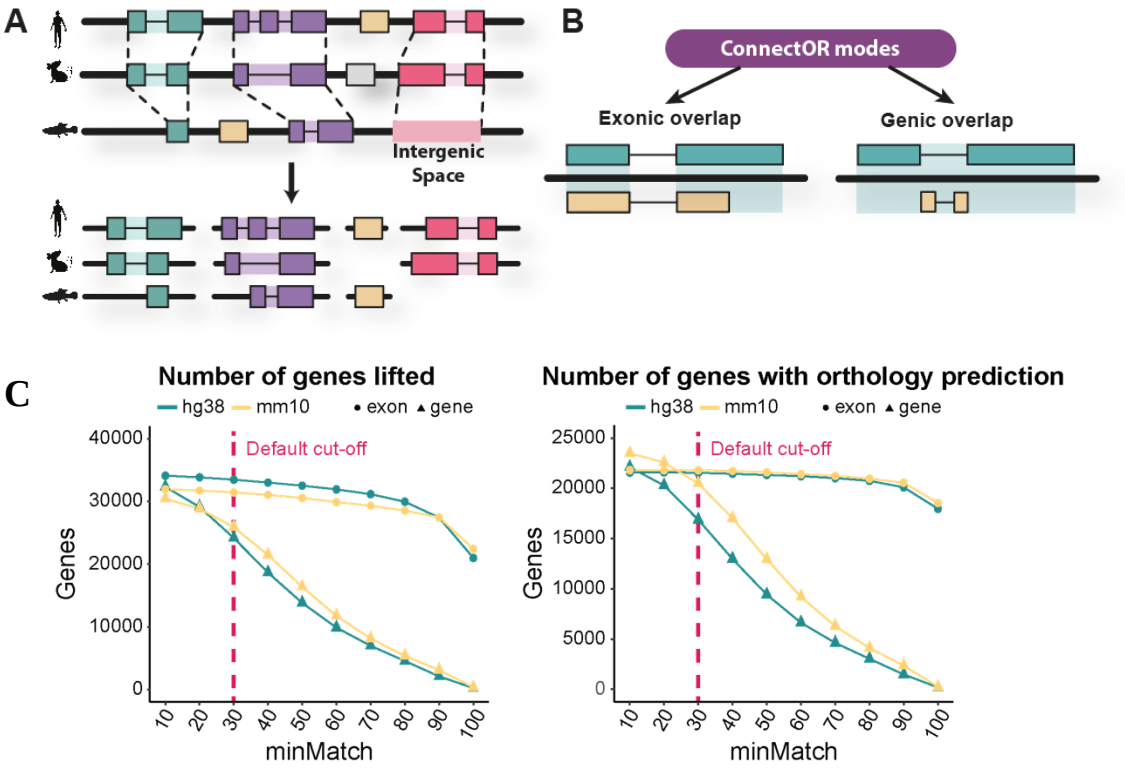


Figure 4.23. Synteny-based approach for lncRNA orthology prediction. *(A) Illustration of conserved syntenic blocks across three species: human, mouse, and zebrafish. (B) Depiction of the two levels of restriction used for orthology prediction: exonic span and genic span. (C) Total number of genes successfully mapped from human to mouse (blue) and from mouse to human (orange) at the exonic span level (circle) and genic span level (triangle) across varying minMatch thresholds. Figure courtesy of prof. Barbara Uszczyńska-Ratajczak.*

The analysis confirmed that, as anticipated, protein-coding genes exhibit a high level of positional conservation, even between distant species. Both mouse and zebrafish share approximately 80% of their genes with humans (Figure 4.24). In contrast, the level of positional conservation for lncRNAs was notably lower, with only around 10% of genes being positionally conserved between humans and mice (Figure 4.24). Additionally, over 60% of orthologous regions fell within intergenic space (Figure 4.24), suggesting the presence of lncRNAs that have yet to be annotated.

Considering that a larger gene catalog enhances the probability and efficiency of ortholog identification, gigaLNC gene set was created by Professor Barbara Uszczyńska-Ratajczak (Figure 4.25). This catalog, by merging all available human lncRNA inventories, represents the largest lncRNA dataset for the human genome.

Using gigaLNC as an input for the analysis significantly improved the identification of positionally conserved orthologs in the mouse genome—by tenfold—and increased the number of regions falling in the intergenic space by approximately sixfold (Figure 4.24). To further improve ortholog detection in zebrafish, the largest available lncRNA gene sets were utilized: gigaLNC (human), NONCODE (mouse), and ZFLNC (zebrafish). This approach identified over 15,000 (~13%) human and 3,000 (~4%) mouse lncRNA orthologues in the zebrafish genome (Figure 4.24). Additionally, about 5,500 human and 2,000 mouse lncRNAs mapped to intergenic regions, suggesting many potentially functional loci yet to be annotated in the zebrafish genome (Figure 4.24).

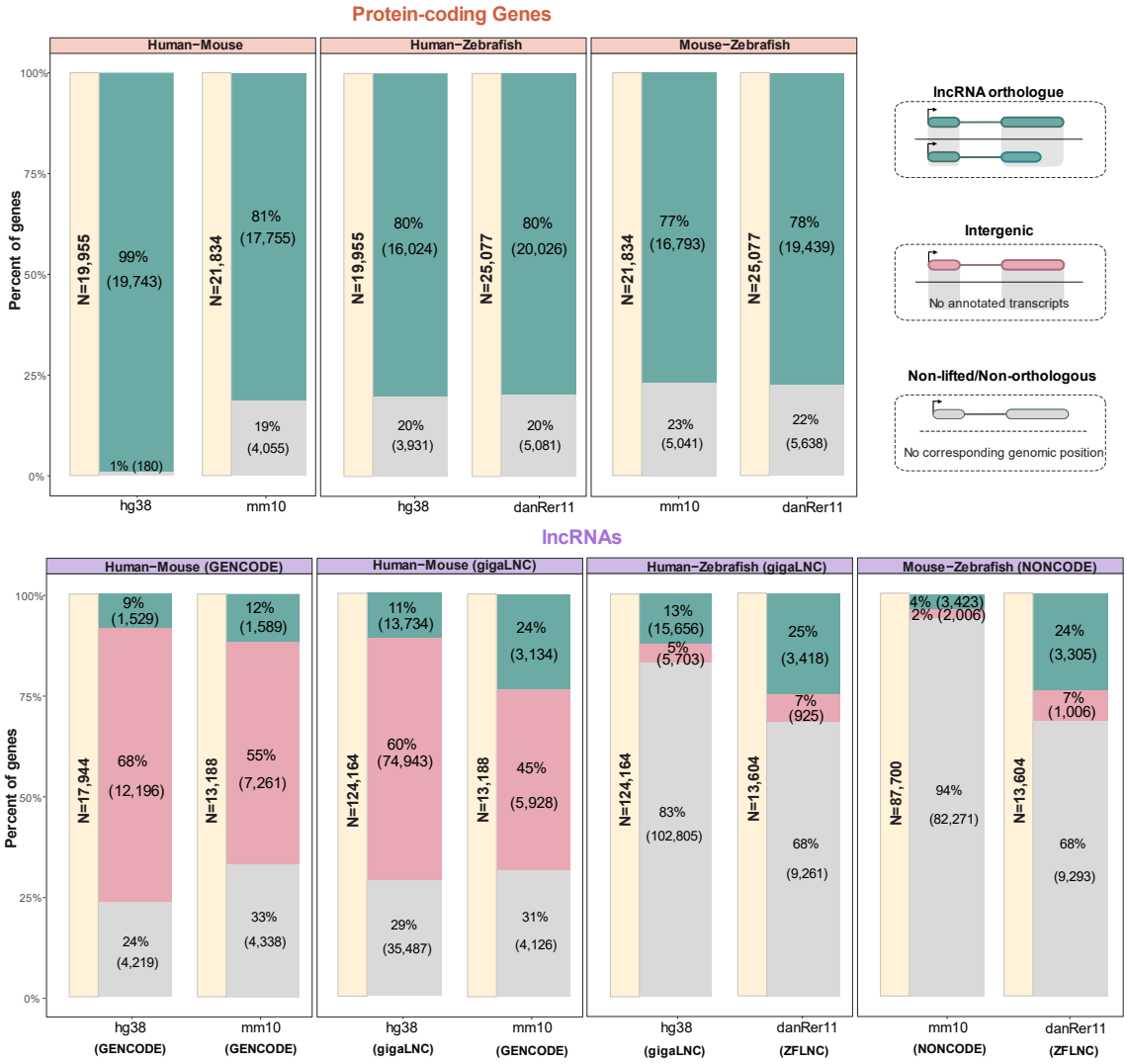


Figure 4.24. Orthology gene prediction between human, mouse and zebrafish. *The number of genes (protein-coding and lncRNA) successfully lifted between human, mouse, and zebrafish using the default minMatch value (30) at the exonic span level. Various gene catalogs were employed for this analysis, including GENCODE and gigaLNC for human, GENCODE and NONCODE for mouse, and ZFLNC for zebrafish. Genes were categorized into three main classes: orthologous (green), intergenic (pink), and non-lifted/non-orthologous (gray).*

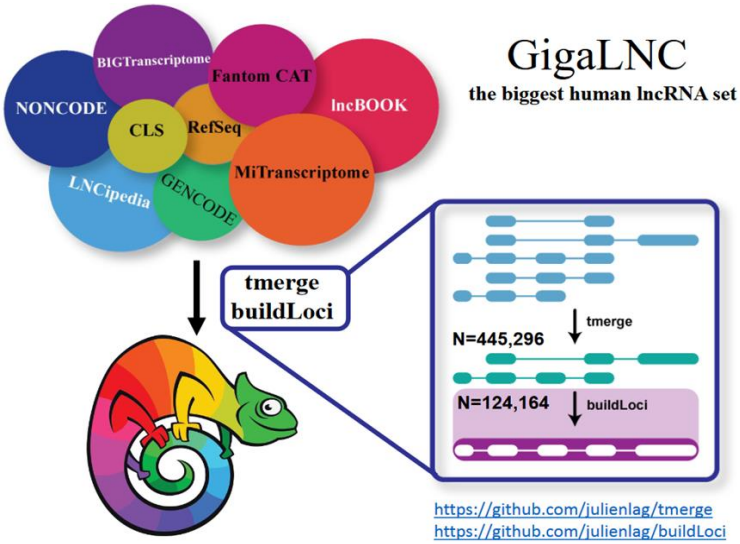


Figure 4.25. Creation process of the biggest catalog of human lncRNAs-gigaLNC. *Figure courtesy of prof. Barbara Uszczyńska-Ratajczak.*

II. Capture probes design for CapTrap-CLS

Please note:
The capture probes were designed by prof. Barbara Uszczyńska-Ratajczak in collaboration with Arbor Biosciences, which provided the customizable myBaits capture solution.

One major challenge in annotating lncRNAs is their low steady-state levels, leading to their underrepresentation in standard, unbiased sequencing libraries. To address this, experimental methods that increase the representation of low-abundance transcripts in cDNA samples are essential. One such method is Capture Long-read Sequencing (CLS), a targeted RNA sequencing technique that combines RNA capture with long-read Oxford Nanopore sequencing (Figure 1.14) (Lagarde et al., 2017). CLS has proven effective in improving lncRNA annotation in human and mouse genomes. Therefore, I decided to leverage this approach to validate its performance in zebrafish.

Using specific probes, CLS can complete existing annotations (green) and identify novel lncRNA loci in suspected regions (purple). In this project, I used CLS to investigate the Ensembl catalog of intergenic lncRNAs, (Known) and thousands of loci that may produce lncRNAs (Novel), such as small RNA genes (miRNAs, snoRNAs and snRNAs) enhancer regions (SE and TE), and ultraconserved noncoding elements (UCNE) (Figure 4.26 A). To further refine the annotation of potentially functional lncRNAs, orthologous lncRNA predictions and intergenic regions obtained from ConnectOR analysis (gigaLNC-human and NONCODE and GENCODE-mouse) (Figure 4.24) were incorporated into probe design.

Next, these candidate regions were filtered to exclude those located within 5 kb of protein-coding genes and those with high expression levels based on publicly available RNA-seq data (Figure 4.26 B, Figure S6), resulting in a refined set of targeted regions (Figure 4.26 A). Altogether, nearly 10 megabases of the zebrafish genome were targeted. As positive controls for enrichment, a set of probes targeting half of the lowly expressed ERCC synthetic spike-ins was included.

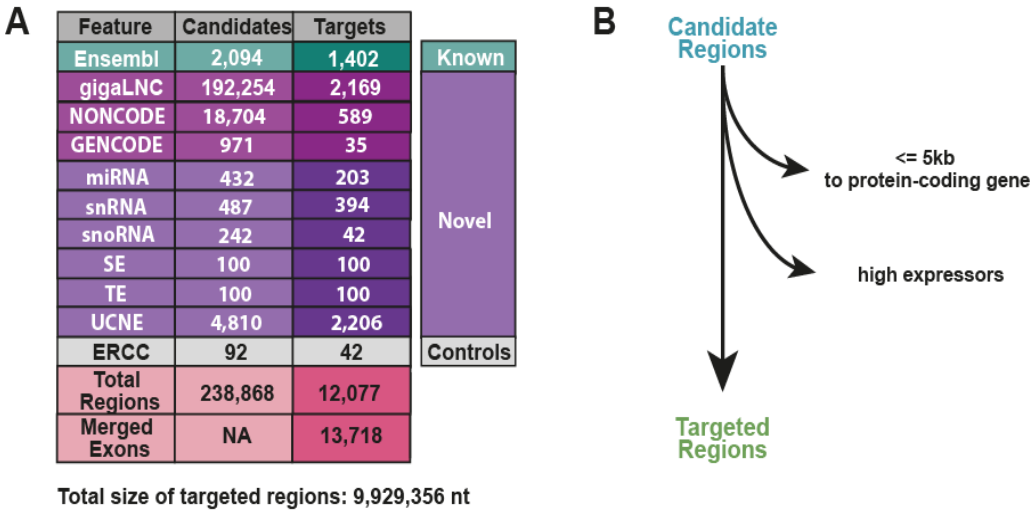


Figure 4.26. Representation of zebrafish capture probe design. *A) List of candidate regions and final targeted regions with numbers indicating the number of gene loci that is going to be probed. Ensembl: intergenic lncRNAs from Ensembl annotation; gigaLNC: positionally conserved regions between human and zebrafish; NONCODE and GENCODE: Positionally conserved regions between mouse and zebrafish; SE: Super-enhancers; TE: Typical enhancers; UCNE: ultraconserved noncoding elements; ERCC: external spike-in sequences. B) Strategy of obtaining targets from candidate regions.*

III. Preparation and Quality Assessment of CapTrap-CLS Libraries

To ensure diversity and reduce the bias of zebrafish lincRNA annotations toward embryonic samples, we focused on transcriptionally complex and biomedically relevant organs, including the brain, heart, liver, testes, and ovaries. Given the known limitations in the quality of current annotations, I also chose to include three zebrafish developmental stages: 2-4 Cell, Shield, and 28 hpf. Based on previous experiments indicating that CapTrap-seq demonstrated optimal performance when combined with the size-selection step (SS500), I decided to integrate it with the Capture Long-read Sequencing (CLS) approach.

I successfully obtained high-quality post-capture libraries (libraries generated after the cDNA capture step) (Figure S7), which were subsequently subjected to ONT sequencing. To further evaluate the performance of CapTrap-CLS, I sequenced both pre- and post-capture libraries. Post-capture samples demonstrated a higher mapping rate in comparison to pre-capture samples (Figure S8). Furthermore, post-capture libraries exhibited higher median and mean read lengths for most samples, with heart samples showing the highest increase of up to 600 bp (Figure S8). Conversely, at the annotation level, post-capture transcript models were generally slightly shorter than their pre-capture counterparts (Figure S9, Figure S10). Post-capture transcript models exhibited a similar proportion of spliced and unspliced transcript models; however, they represented a slightly lower fraction of 5' and 3'-complete transcript models compared to the pre-capture samples (Figure S11). These results are consistent with observations from the performance analysis of CapTrap-CLS in human and mouse samples (Kaur et al., 2024). The assessment of transcript 5'-end completeness was performed by evaluating their proximity to CAGE tags provided by DANIO-CODE. Therefore, the lower fraction of complete transcripts is not surprising, as CAGE profiling of transcription start sites (TSS), like other RNA-seq methods, is highly dependent on expression levels. Furthermore, as previously mentioned, zebrafish annotation studies have primarily relied on shallow RNA-seq data. Consequently, identifying novel TSS for lowly expressed lncRNAs remains considerably more challenging than for protein-coding genes, which increases the risk of false negative conclusions regarding their 5'-end completeness.

IV. CapTrap-CLS performance in zebrafish

The performance of CapTrap-CLS was assessed at multiple levels. Initially, I examined the enrichment effect on the positive ERCC controls targeted by capture probes. This evaluation employed two metrics: the "on-target" rate, which reflects the proportion of reads derived from targeted regions, and enrichment, indicating the increase in the on-target rate following capture. I observed a 40-fold increase in reads mapping to the targeted ERCCs in the post-capture libraries. Specifically, over 95% of the reads mapping to ERCCs were attributed to the targeted probes, compared to only 3% in the pre-capture samples (Figure 4.27). The results were highly comparable to those obtained from CapTrap-CLS applied to human samples (Figure 4.27), as well as to those from mouse samples (Kaur et al., 2024).

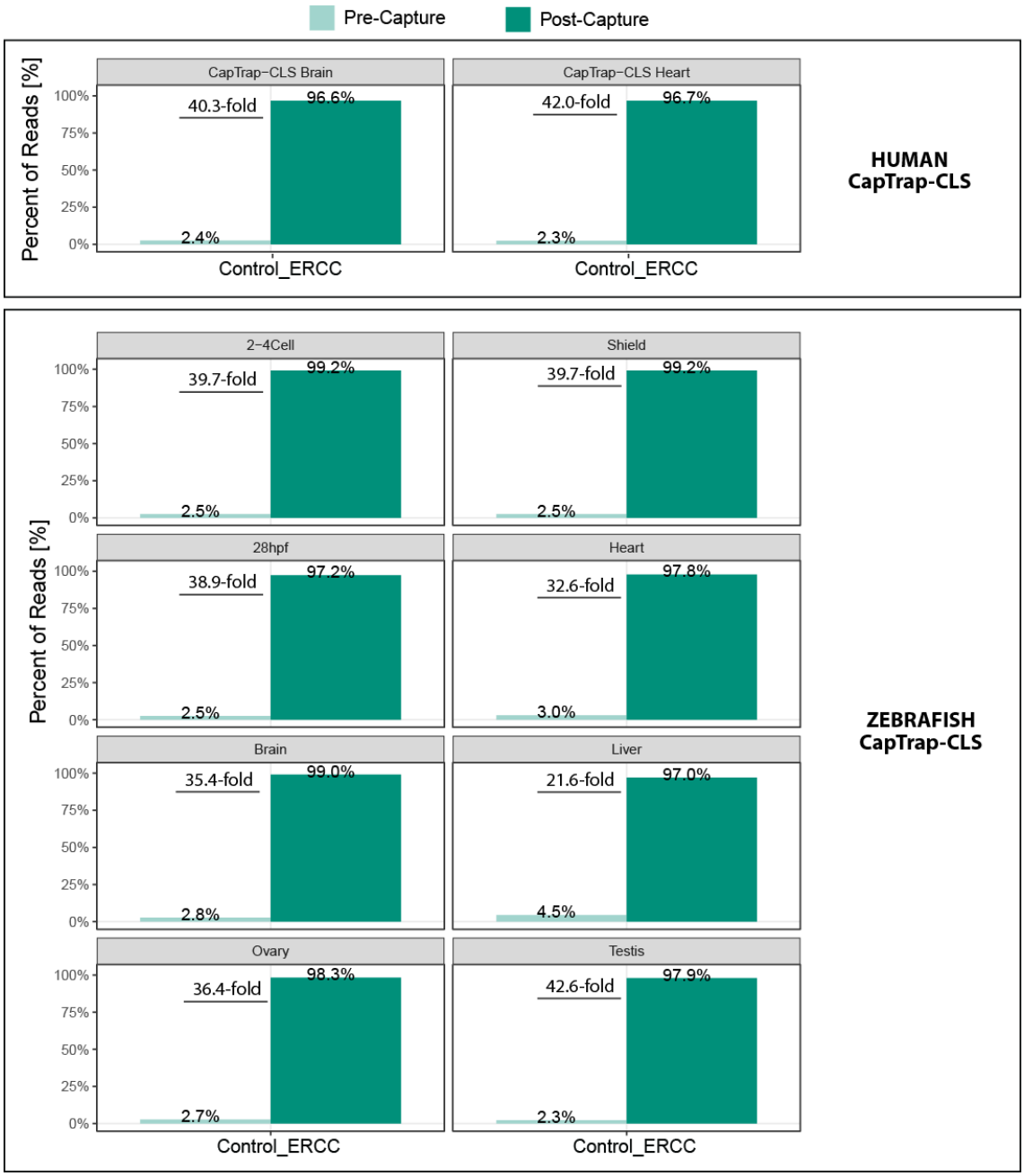


Figure 4.27. cDNA capture performance. *The y-axis shows the percentage of all mapped RNA-seq reads originating from a targeted regions.*

Next, I examined how capture steps impacted the targeting of ENSEMBL gene biotypes. The analysis revealed a 10-fold increase in reads mapping to lncRNA genes in post-capture samples. Additionally, other categories that showed significant increases included non-exonic regions, with the highest fraction observed in the testis; pseudogenes, which had the largest fraction in the liver; and poorly characterized miscellaneous RNA in the shield sample (Figure 4.28).

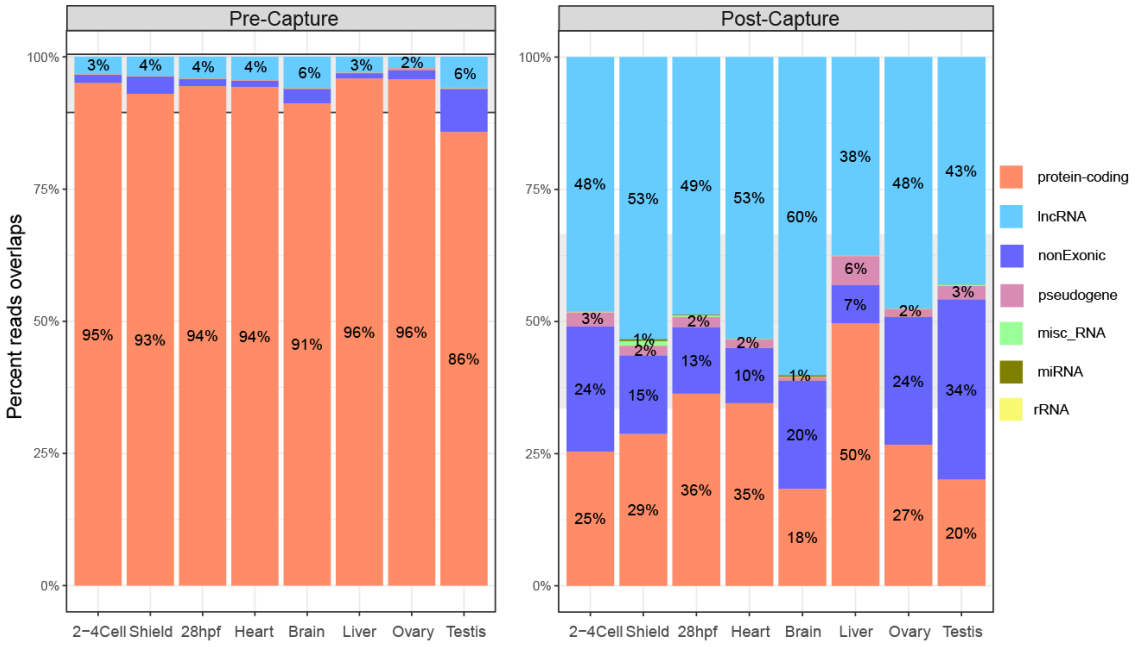


Figure 4.28. GENCODE gene biotype detection in pre- and post-capture libraries. *The stacked bar plots show the percentage of raw reads mapping to various annotated ENSEMBL gene biotypes. Different colors represent the ENSEMBL gene biotypes: protein-coding (orange), ribosomal RNAs (yellow), pseudogenes (pink), miscellaneous RNA (green), miRNAs (olive green), non-exonic (dark blue), and lncRNAs (light blue).*

Subsequently, I conducted a more detailed examination of the post-capture effect on targeted regions, as well as the fraction of reads located in intergenic regions. The CapTrap-CLS approach achieved on-target rates ranging from 47.1% to 79.0%, with enrichment levels varying from 18-fold to exceeding 100-fold, depending on the tissue analyzed (Figure 4.29). These results significantly surpass the original CLS performance metrics for human (29.7%) and mouse (16.5%) samples. The CapTrap-CLS method in human and mouse samples resulted in 5- to 35-fold enrichment of targeted regions (Kaur et al., 2024), further indicating a more efficient enrichment in zebrafish across a range of tissues. These differences can be attributed to several factors, including the type of probes used for cDNA capture, the implementation of an additional size-selection step, or the size of the targeted regions themselves. However, further testing is required to assess the relative contribution of each of these elements.

The implementation of the CapTrap-CLS approach facilitated the identification of novel loci across all tested samples. Notably, in testis samples, over 30% of the reads mapped to intergenic regions, highlighting its efficacy in uncovering new genomic elements (Figure 4.29). Similarly, the application of CapTrap-CLS to human testis samples yielded the highest level of novelty, as reported by Kaur et al., 2024, further emphasizing its utility in transcriptome exploration and annotation.

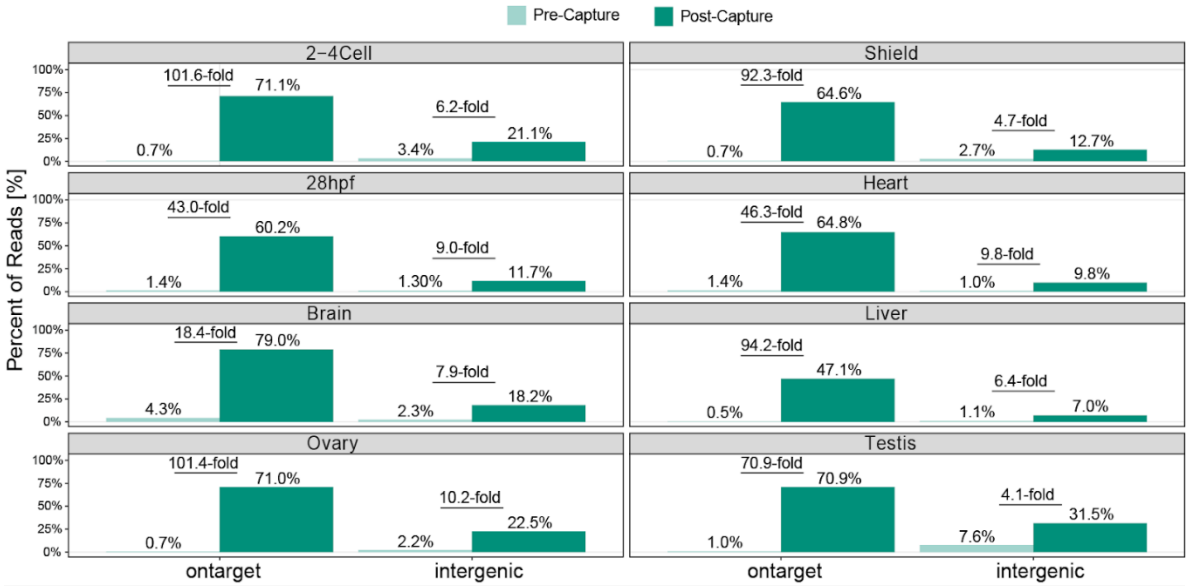
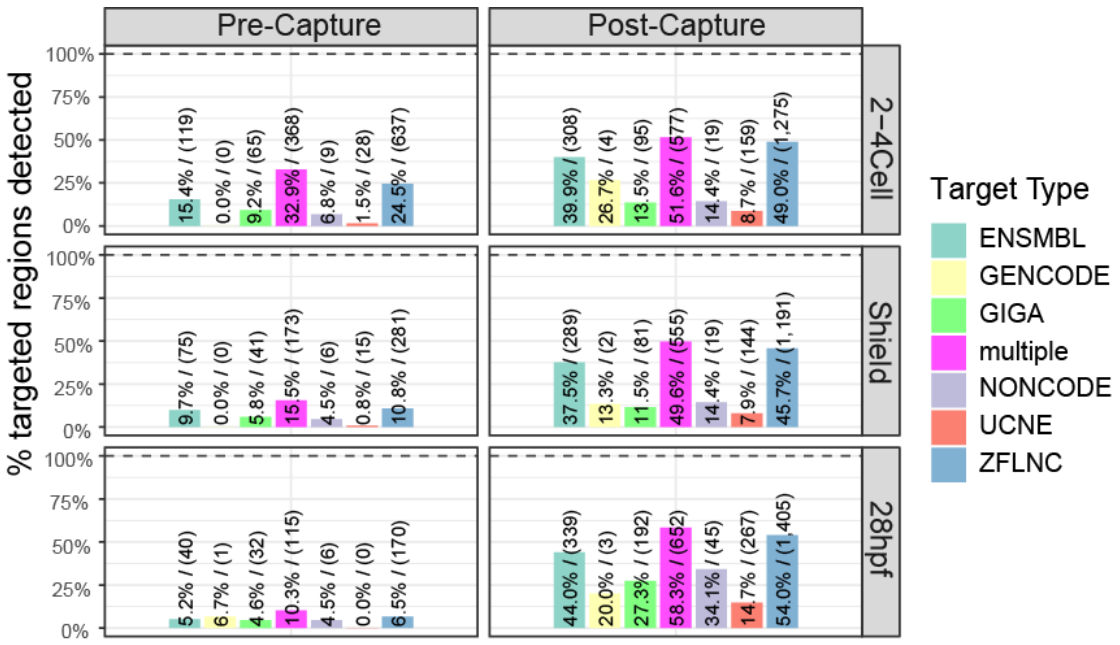


Figure 4.29. cDNA capture performance. *The y-axis shows the percentage of all mapped RNA-seq reads originating from a targeted regions.*

Furthermore, I successfully identified in the sequencing library approximately 20% of the regions predicted to be positionally conserved between humans and zebrafish, and around 30% between mice and zebrafish. The highest detection rates in the post-capture samples were observed for ENSEMBL and ZFLC targets categories. Conversely, the lowest detection rate was recorded for Ultraconserved Noncoding Elements (UCNE); however, this category exhibited the most significant fold change in detection between pre- and post-capture samples. The most notable change between pre- and post-capture samples was observed in the liver (Figure 4.30).



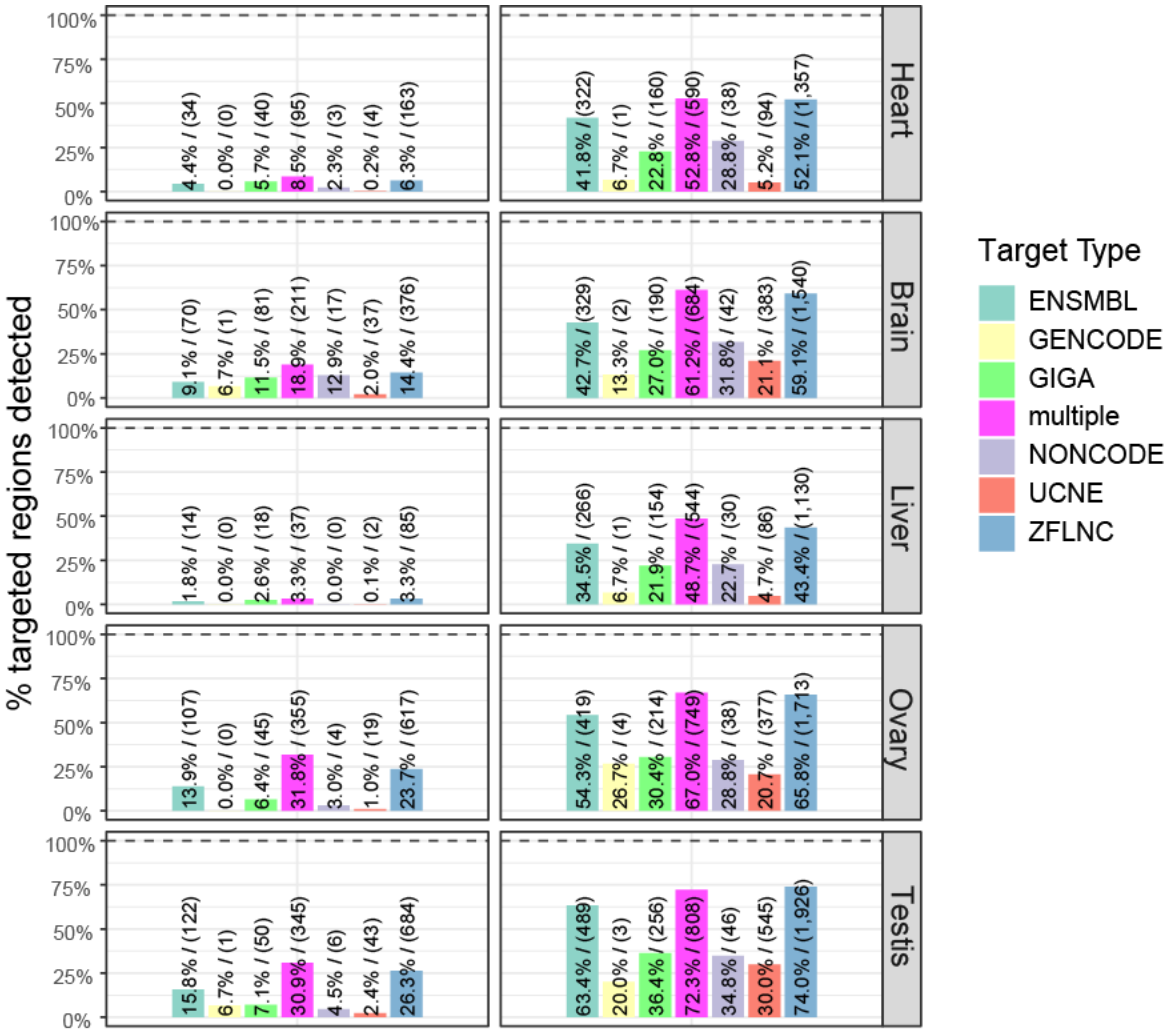


Figure 4.30. The detection rate of selected targets in pre- and post-capture across zebrafish samples *is illustrated, with each target category represented by a distinct color. The count of detected targets within each category is displayed above each bar, accompanied by the corresponding percentage.*

V. The Impact of CapTrap-CLS on ENSEMBL annotation extension

To evaluate the impact of CapTrap-CLS on extending Ensembl lncRNA annotation, coding potential analyses of transcript models generated in this study were conducted using CPC2 (Kang et al., 2017) and CPAT (Wang et al., 2013). Each tool categorized transcripts as either non-coding RNAs or protein-coding. To create a definitive set, transcripts predicted as non-coding by CPC2 were intersected with CPAT non-coding predictions, extracting common models. A similar approach was applied for protein-coding predictions. This process yielded two robust sets: one for lncRNAs and another for protein-coding RNAs.

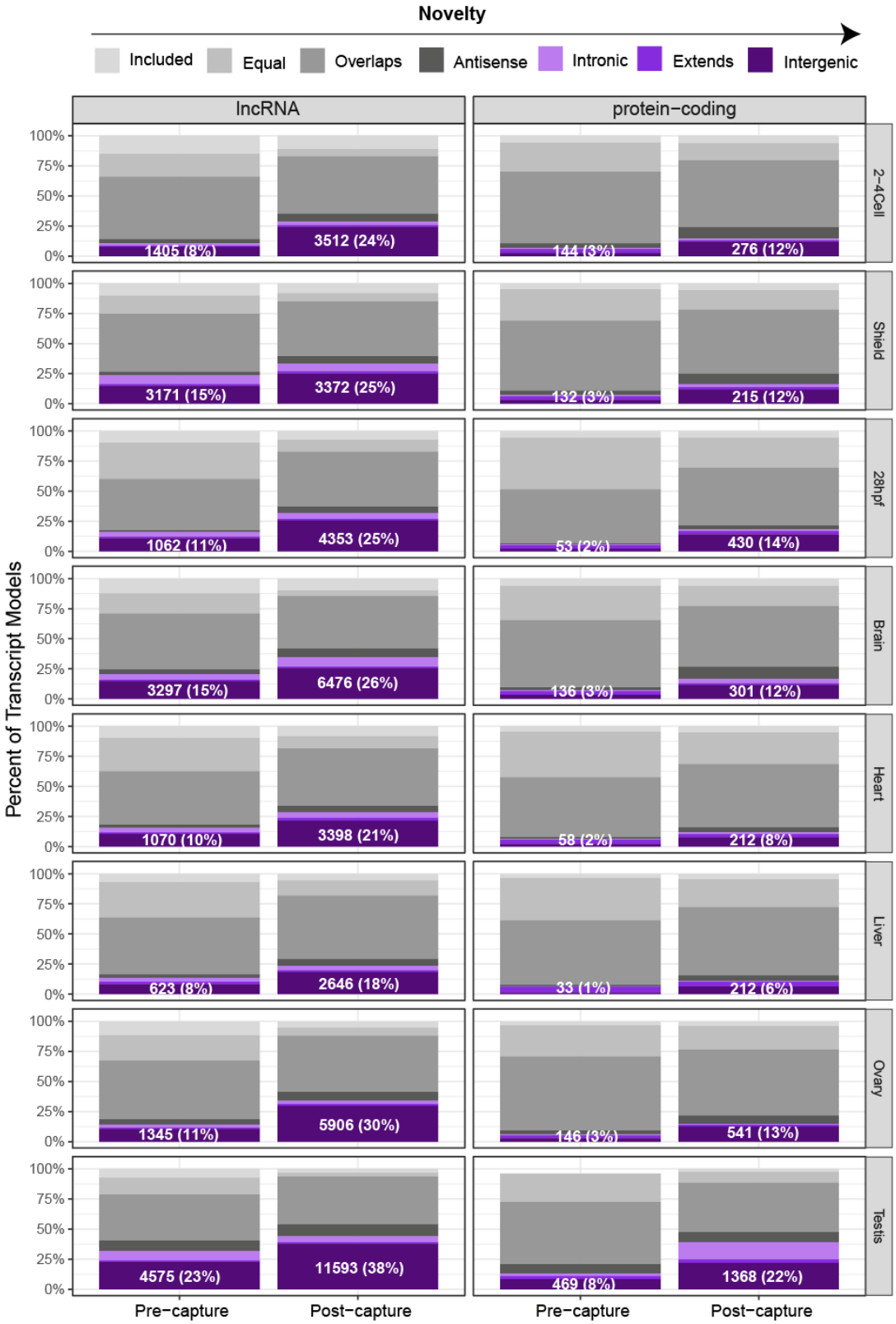


Figure 4.31. Impact of the CapTrap-CLS approach on extending ENSEMBL annotation. *Intersected sets of non-coding and protein-coding transcript models (TMs), derived from coding potential analyses, were compared against the ENSEMBL reference. TMs were categorized into seven classes: Included, Equal, Overlaps, Antisense, Intronic, Extends, and Intergenic. Deeper shades of purple in the classification indicate higher levels of novelty in the transcript models.*

The intersected sets of non-coding and protein-coding transcripts were then compared with ENSEMBL annotations to evaluate the extension of annotation, with a particular focus on intergenic (novel) transcript models. As anticipated, the CapTrap-CLS approach had a greater impact on identifying novel non-coding RNAs compared to mRNAs (Figure 4.31). The proportion of intergenic transcript models increased from 10-15% in pre-capture samples to approximately 25% in post-capture samples (Figure 4.31).

Testis samples exhibited the highest overall identification of novel transcript models. However, the greatest fold increase in the number of novel loci between pre- and post-capture samples was observed in ovary, liver, and 28 hpf samples, with these categories showing more than a fourfold rise in novel loci detection (Figure 4.32).

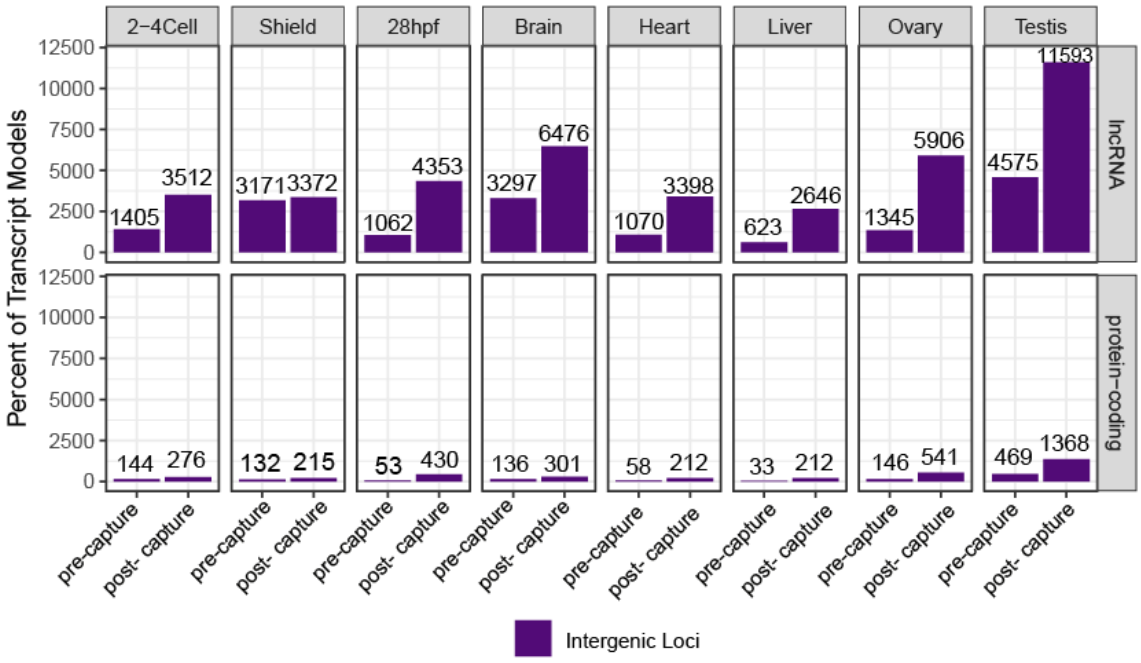


Figure 4.32. The number of intergenic transcripts identified in pre- and post-capture samples across various developmental stages and adult tissues *was assessed, excluding all loci overlapping the Ensembl v104 annotation.*

The integration of full-length library preparation and long-read sequencing significantly enhanced the zebrafish Ensembl annotation. A comparison of the whole annotation created in this study with the Ensembl v104 catalog revealed over 14,000 novel genes and more than 200,000 transcripts, resulting in a nearly fourfold increase in the mean isoform count per gene (Ensembl++) (Figure 4.33).

The implementation of CPC2 and CPAT protein-coding potential prediction tools helped identify that the vast majority of these novel loci are predicted to be lncRNAs. These newly identified lncRNA loci contributed significantly to the creation of the Ensembl+ catalog, which was generated by merging the Ensembl annotation with intersected non-coding RNA annotation set. This effort resulted in the addition of 12,000 novel lncRNA loci and achieved a twofold increase in the average isoform count per gene for this RNA type. This marks the greatest improvement in zebrafish lncRNA annotation to date (Figure 4.33).

A significant contribution to this advancement came from the application of the CapTrap-CLS approach, which alone enabled the detection of nearly 25,000 genes predicted to be lncRNAs. The CapTrap-CLS lncRNA catalog stands out due to its high fraction of complete transcript models—60%—compared to around 30% for GENCODE v47 lncRNAs. Additionally, it shows a higher mean number of isoforms per gene (8 for CapTrap-CLS in zebrafish versus 6 for GENCODE v47 lncRNAs) (Figure 4.33).

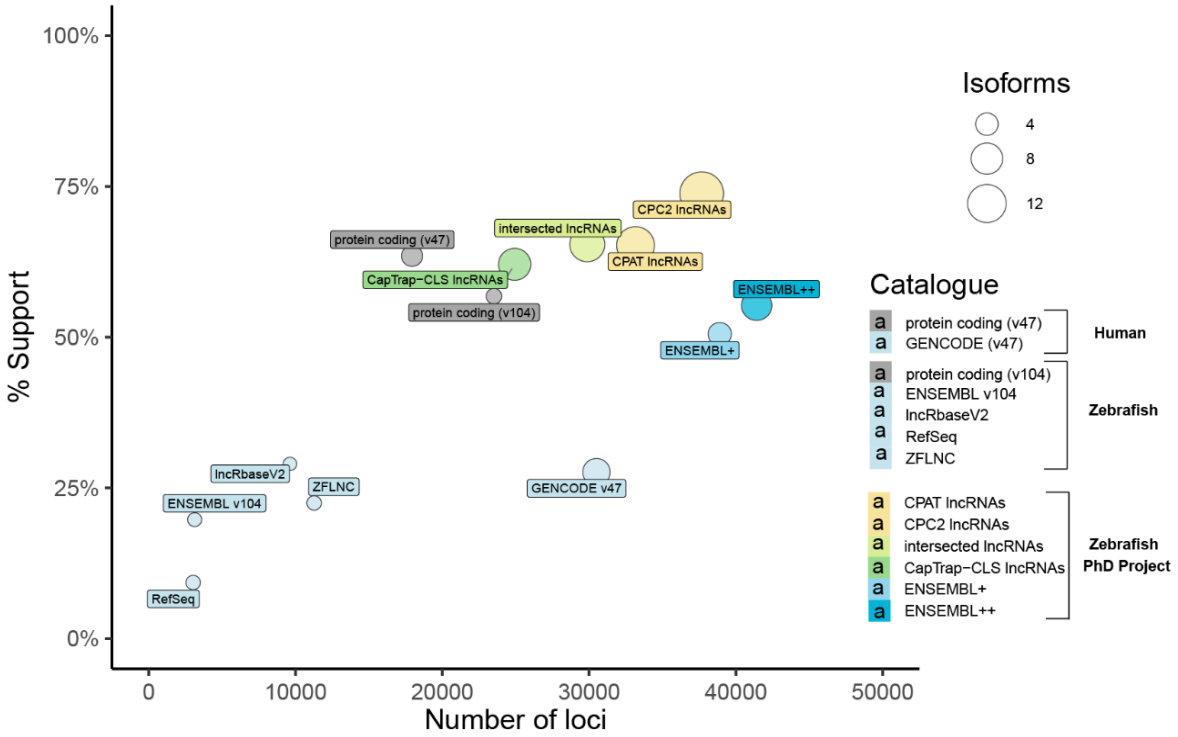


Figure 4.33. Comparison of quality metrics across zebrafish lncRNA annotations and catalogs created in this study. *If the genome build of analyzed annotation predates GRCz11, it was lifted to this version. To ensure consistency, assembly patches were excluded, and gene loci boundaries were redefined using buildLoci (https://github.com/julienlag/buildLoci), resulting in expected differences in gene locus counts compared to the reported lncRNA in each annotation. The x-axis represents the total number of gene loci in each annotation, while the y-axis indicates the percentage of transcript structures supported by a CAGE cluster and polyA signal. The diameters of the circles represent the mean number of transcripts per gene. The "protein-coding" and "GENCODE" (v47) categories represent confidently annotated GENCODE protein-coding and lncRNA genes for the human genome, respectively (Uszczyńska-Ratajczak, 2018, Kaur et al., 2024). Protein-coding catalogs are shown in gray, while lncRNA catalogs are shown in blue. CPC2 and CPAT lncRNAs: annotation created from CPC2 and CPAT noncoding TMs sets respectively from all biological samples and library preparation protocols tested; Intersected lncRNAs: set of lncRNAs overlapping both CPC2 and CPAT noncoding RNAs; CapTrap-CLS lncRNAs: annotation created from CapTrap-CLS samples; ENSEMBL+: annotation created by merging the whole ENSEMBL v104 annotation with intersected lncRNAs set; ENSEMBL++ annotation created by merging the whole ENSEMBL v104 with total transcript models set generated in this study, all samples.*

PART C. Functional characterization of identified lncRNAs

The final stage of my PhD project focused on evaluating whether positional conservation can reliably predict the biological function of lncRNAs. Additionally, I aimed to assess the impact of the extended lncRNA annotation for zebrafish, generated through CapTrap-CLS, on the functional characterization of selected targets.

Initially, the ConnectOR (https://github.com/Carlospq/ConnectOR) synteny-based approach was applied using reference annotations from human (GENCODE) and zebrafish (Ensembl) to identify positionally conserved lncRNAs. Out of more than 300 orthologous lncRNAs, only lncRNA-to-lncRNA predictions were selected (49 genes) (Figure 4.34). The 49 genes were then thoroughly analyzed to identify the most promising candidates for functional validation studies.

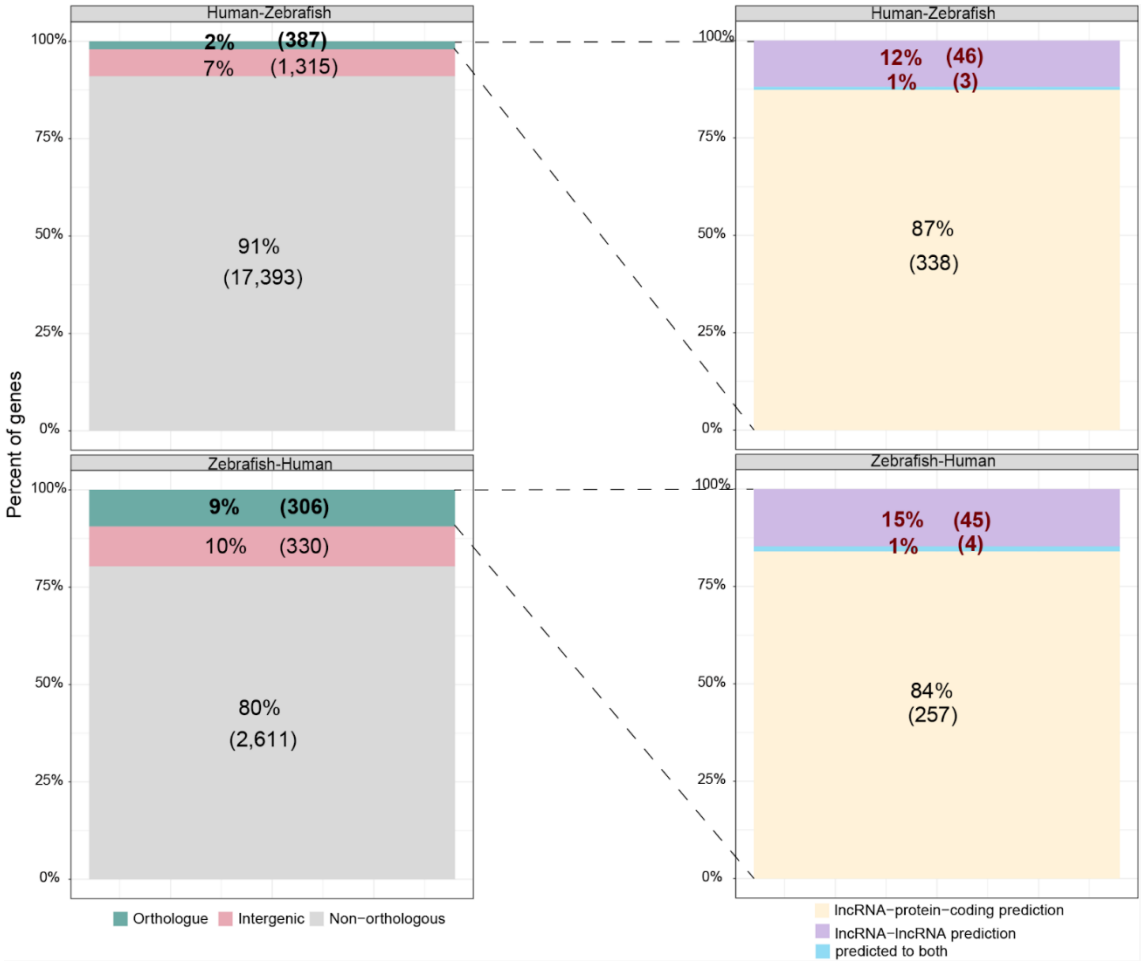


Figure 4.34. Positionally conserved lncRNAs between human (GENCODE) and zebrafish (Ensembl). *Genes were categorized into three main classes: orthologous (green), intergenic (pink), and non-lifted/non-orthologous (gray).Orthologous genes found with ConnectOR were further classified based on conserved gene biotype: lncRNA-to-lncRNA prediction (purple), lncRNA-to-protein-coding prediction (yellow) and predicted to both (blue). LncRNAs selected for further analysis lncRNA-to-lncRNA predicted were marked with burgundy font.*

Positionally conserved lncRNAs were subsequently analyzed for functional associations, including validated links to diseases (EvLnc2; Zhou et al., 2021) and cancer (CLC3; Vancura et al., 2022). The analysis revealed that 16 positionally conserved lncRNAs were associated with diseases, while 17 were linked to cancer development (Figure 4.35). Notably, a significant proportion of the positionally conserved lncRNAs were identified as small RNA host genes, with 10 out of the 49 categorized as miRNA host genes (MIRHGs) and 1 classified as a snoRNA host gene (SNHG). Furthermore, more than half of these small RNA host genes (6 out of 11) were found to be associated with disease or cancer progression (Figure 4.35).

Figure 4.35. The associations of positionally conserved lncRNAs with diseases and cancer. *LncRNAs are classified into different categories, each represented by distinct colors. Furthermore, the names of snoRNA host genes (SNHGs) were highlighted in purple font and miRNA host genes (MIRHGs) highlighted in green font.*
Data is provided courtesy of Dr. Daniel Kuźnicki.

Furthermore, it can be hypothesized that elevated expression of certain genes during specific developmental stages or within particular tissues may signal their biological significance. To investigate the expression patterns of positionally conserved lncRNAs, I analyzed publicly available RNA-seq data covering various zebrafish developmental stages (Pauli et al., 2012). This analysis revealed that most positionally conserved lncRNAs had low expression levels throughout zebrafish development. However, a subset, particularly small RNA host genes, exhibited elevated and stage-specific expression patterns. For example, *ENSDARG00000103682*, the zebrafish ortholog of the human *SNHG1* gene, showed the highest expression among all positionally conserved lncRNAs. Additionally, miRNA host genes often displayed increased expression between 1 days post fertilization (dpf) and 2 dpf (Figure 4.36), a critical period during which key organ development processes, such as heart and brain formation, occur in zebrafish.

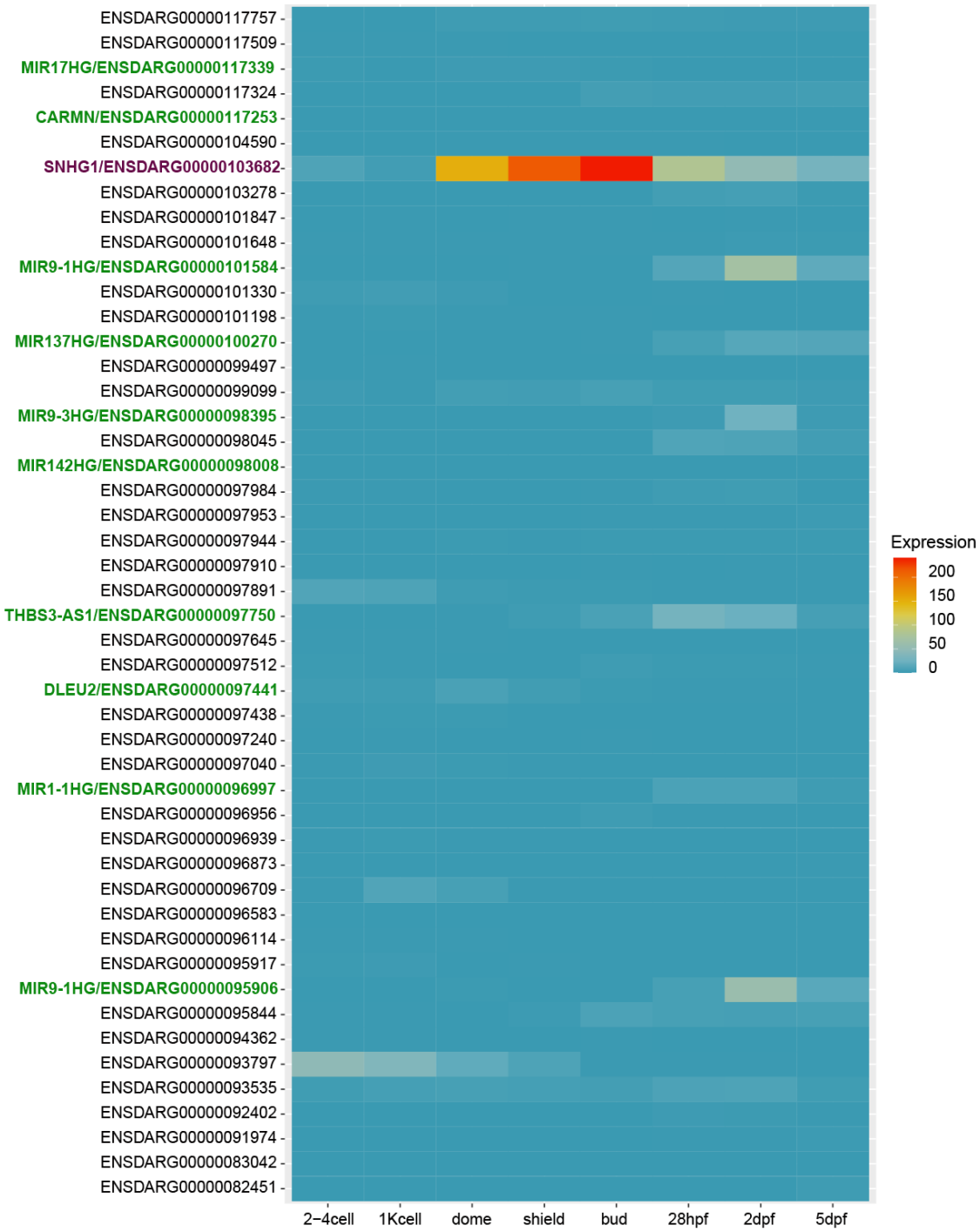


Figure 4.36. The expression patterns of 49 positionally conserved lncRNAs during zebrafish development. *The name of the snoRNA host gene (SNHG) was highlighted in purple, alongside the corresponding lncRNA gene names in the human genome. Similarly, miRNA host genes (MIRHGs) were highlighted in green.*

This observation prompted me to investigate the tissue-specific expression of small RNA host genes in adult zebrafish individuals. To explore this, I utilized publicly available RNA-seq data coming from a panel of adult organs and tissues (Kaushik et al. 2013). Small RNA host genes were found to exhibit tissue-specific expression, with six miRNA host genes specifically expressed in the brain, three in the heart, and one in the blood (Figure 4.37). Notably, most brain-specific miRNA host genes demonstrated higher expression levels compared to those expressed in the heart. In contrast, *snhg1* displayed the highest expression among all small RNA host genes, particularly being expressed in the liver (Figure 4.37).

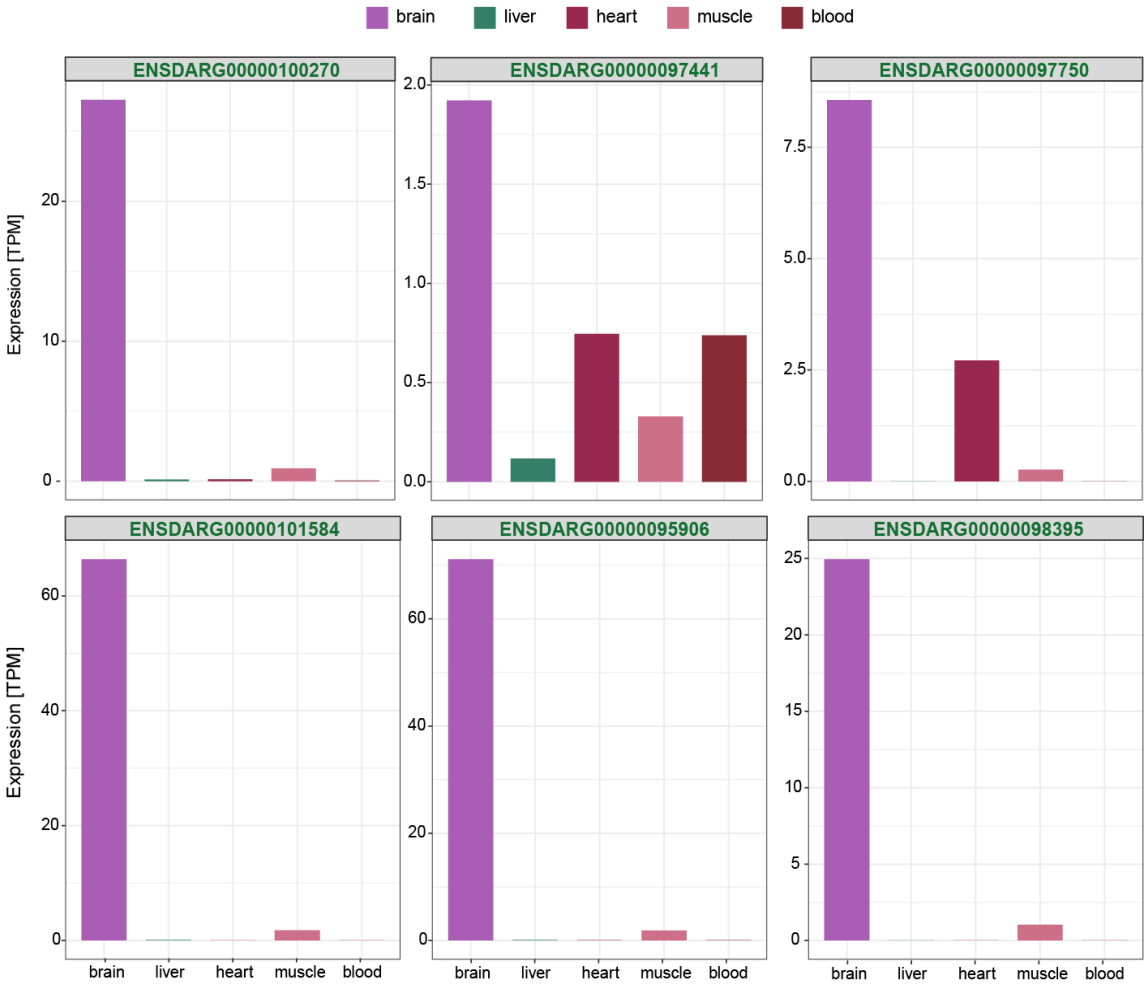


Figure 4.37. Expression patterns of positionally conserved small RNA host genes in adult zebrafish tissues. *Each tissue is represented by a different color. The names of small RNA host genes are displayed above each graph in bold. SnoRNA host gene name is highlighted in purple font, while miRNA host genes are represented in green font.*

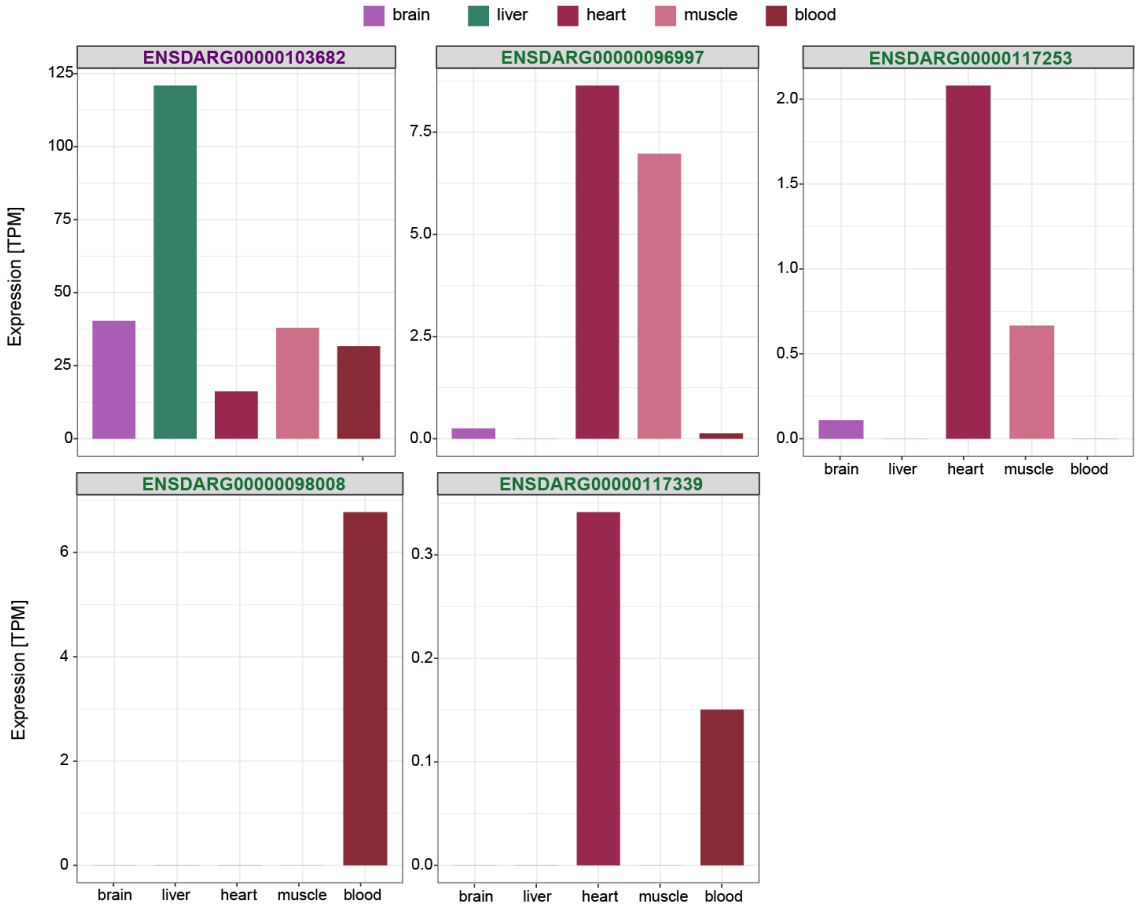


Figure 4.37. Expression patterns of positionally conserved small RNA host genes in adult zebrafish tissues. *Each tissue is represented by a different color. The names of small RNA host genes are displayed above each graph in bold. SnoRNA host gene name is highlighted in purple font, while miRNA host genes are represented in green font.*

Next, I investigated the impact of CapTrap-CLS application on annotation extension of these small RNA host genes. miRNAs and snoRNAs can be encoded within exonic or intronic regions of lncRNAs or protein-coding genes, meaning they are co-transcribed as part of their host genes (Sun et al, 2021; Monziani and Ulitsky, 2023). During splicing, these regions containing small RNAs are excised and subsequently processed into their mature RNA forms. Upon investigating the CapTrap-CLS annotation, I observed that positionally conserved small RNA host genes displayed highly complex alternative isoform patterns, which influenced the location of small RNA-encoding regions into exonic or intronic regions of the host gene. miRNA host genes primarily utilized alternative transcription start sites (TSS), alternative transcription termination sites (TTS), or exon skipping to regulate their encoded miRNAs (Figures 4.38; 4.39; 4.40), while for *snhg1*, intron retention events were the predominant mechanism (Figure 4.41).

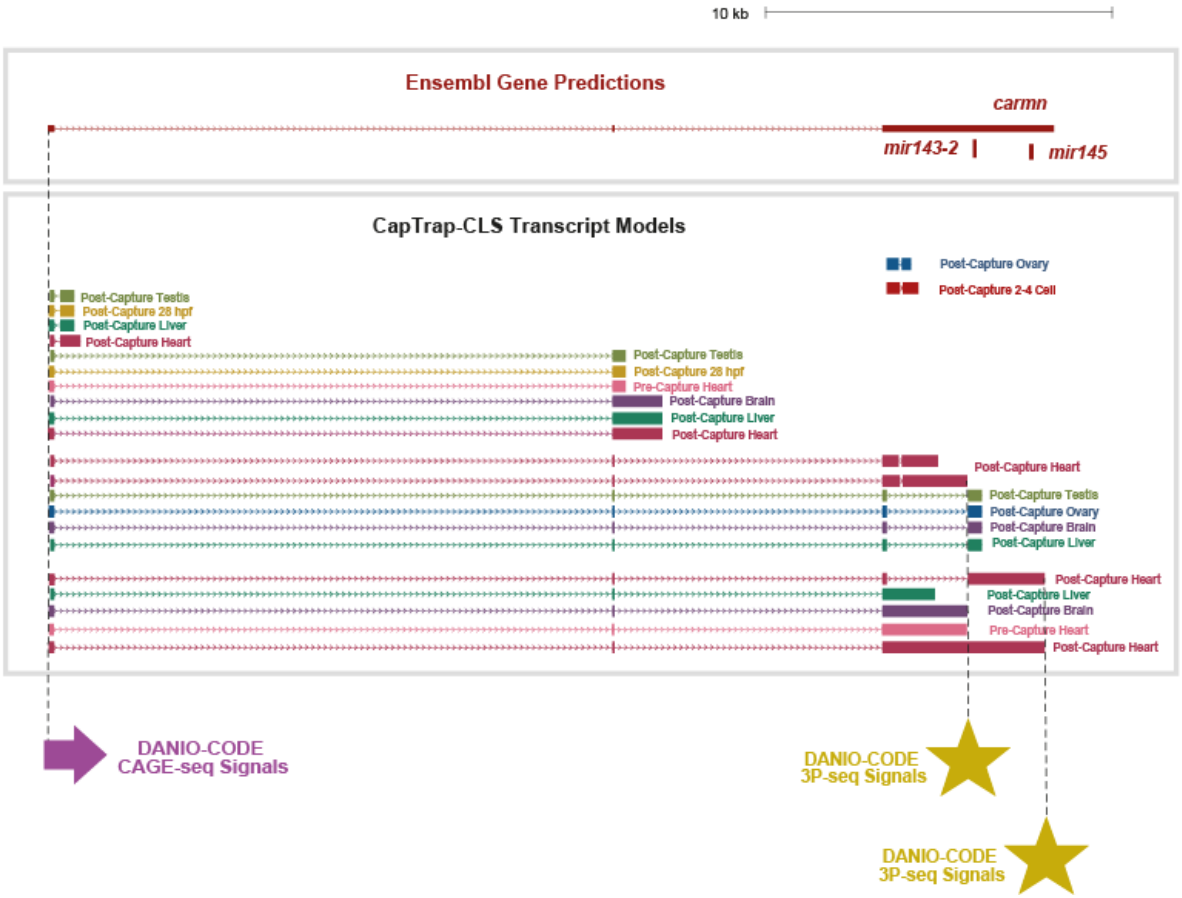


Figure 4.38. Novel transcript isoforms identified for *carmn* loci. *CAGE-seq is indicated by a purple arrow; 3'-end support from 3P-seq is represented by a yellow star.*

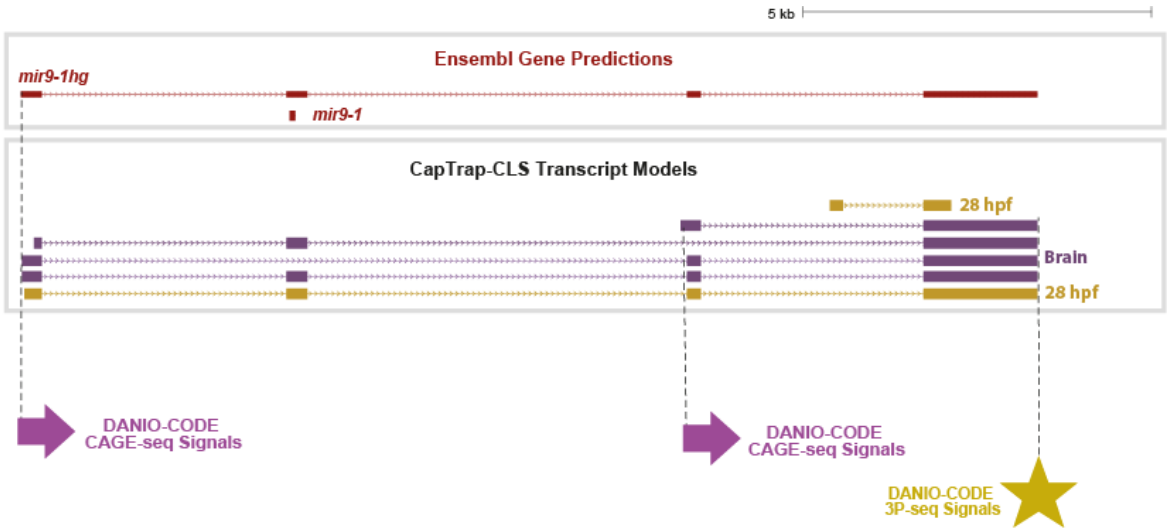


Figure 4.39. Novel transcript isoforms identified for *mir9-1hg* loci. *CAGE-seq is indicated by a purple arrow; 3'-end support from 3P-seq is represented by a yellow star.*

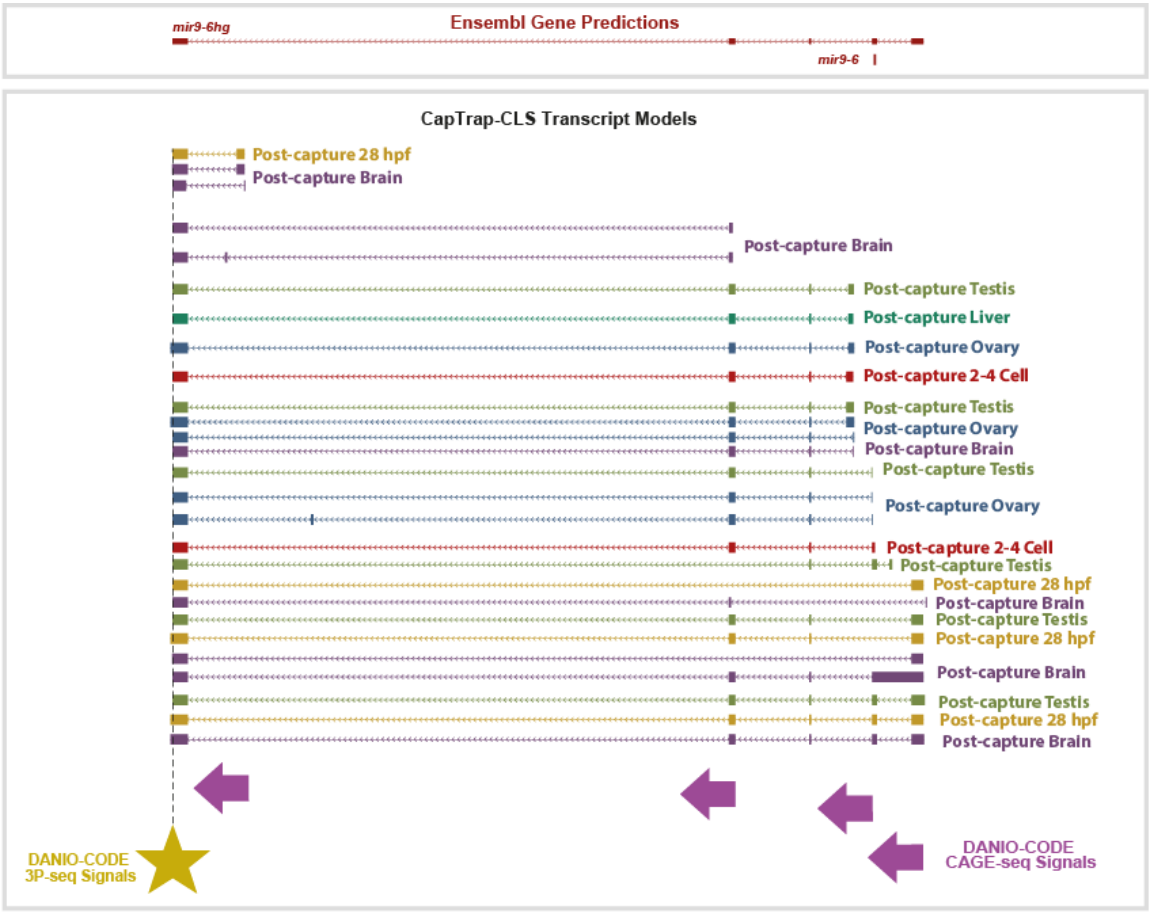


Figure 4.40. Novel transcript isoforms identified for *mir9-6hg* loci. *CAGE-seq is indicated by a purple arrow; 3'-end support from 3P-seq is represented by a yellow star.*

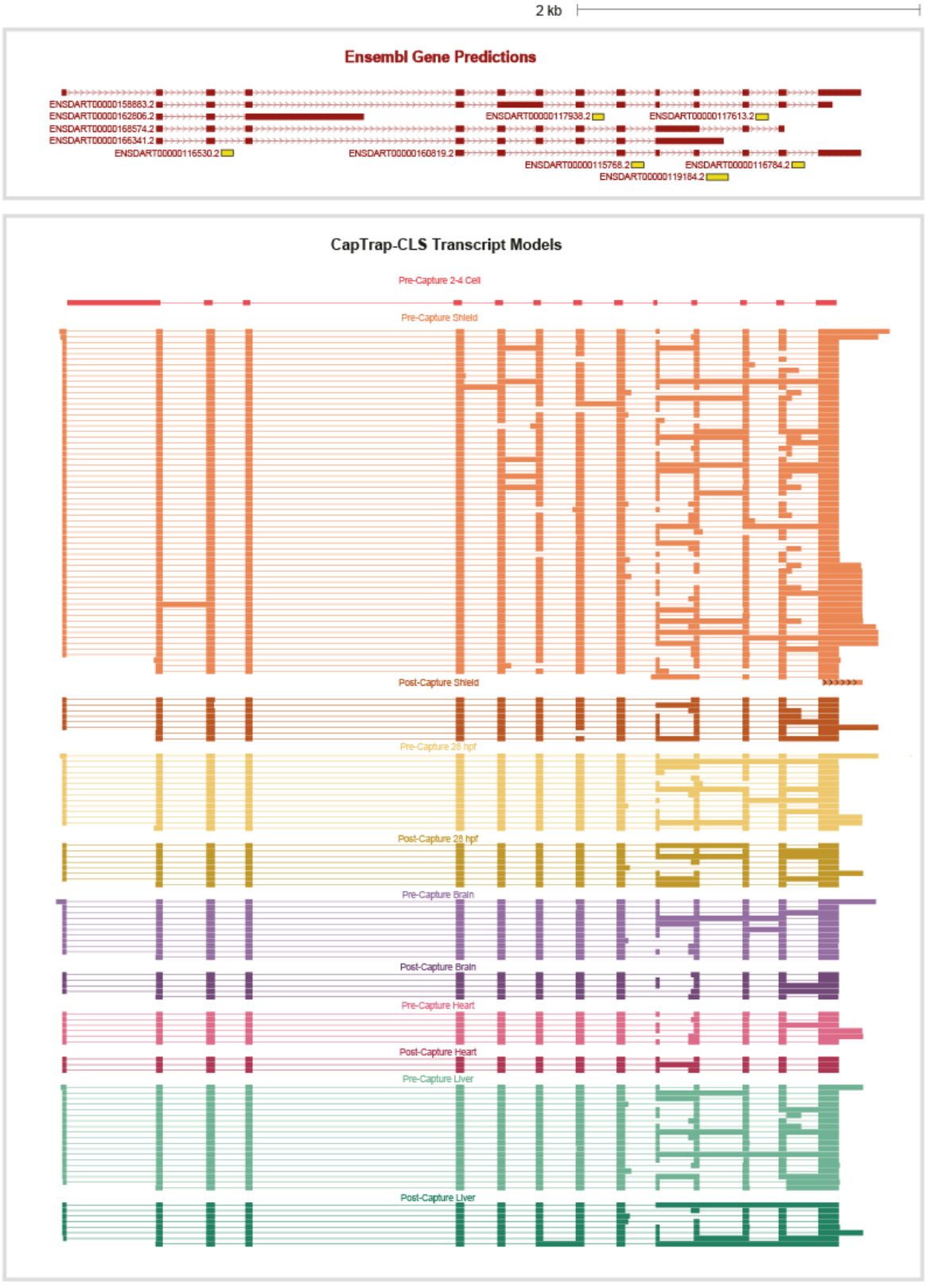


Figure 4.41. Novel transcript isoforms identified for *snhg1* loci. *snoRNAs (ENSEMBL) encoded within the snhg1 gene were presented in yellow.*

For further functional characterization, I selected *snhg1*, which exhibited the highest expression among all positionally conserved lncRNAs. My objective was to investigate its expression patterns during organ development in zebrafish at 1 day post-fertilization (dpf). RNAscope experiments demonstrated that *snhg1* displayed specific expression patterns in the brain (Figure 4.42 and Figure 4.43) and in the eye regions (Figure 4.44).

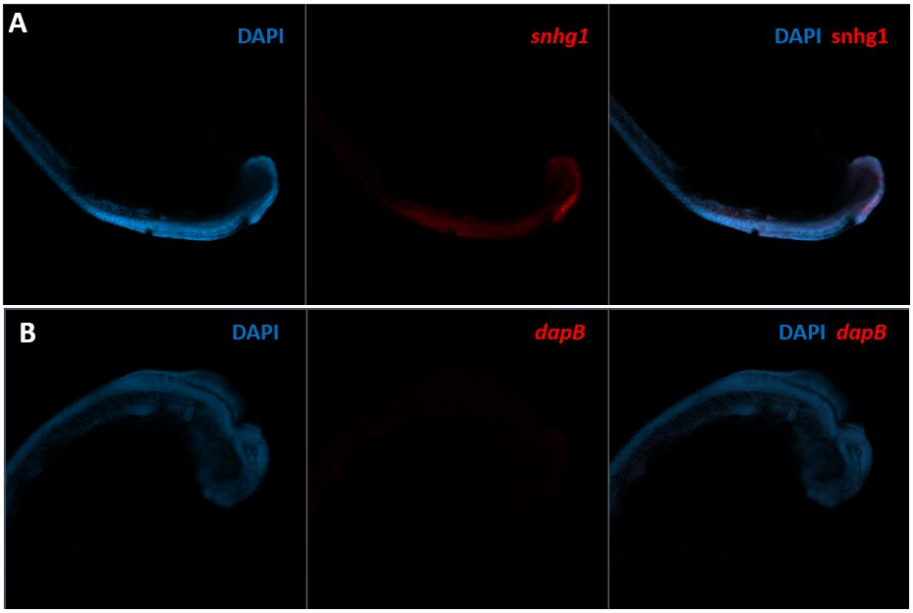


Figure 4.42. Expression profile *snhg1* at 28 hours post fertilization. *A) snhg1 gene and B) dapB-negative control in zebrafish Fluorescent signal coming from RNAscope probes is shown in red, while DAPI signal in blue.*

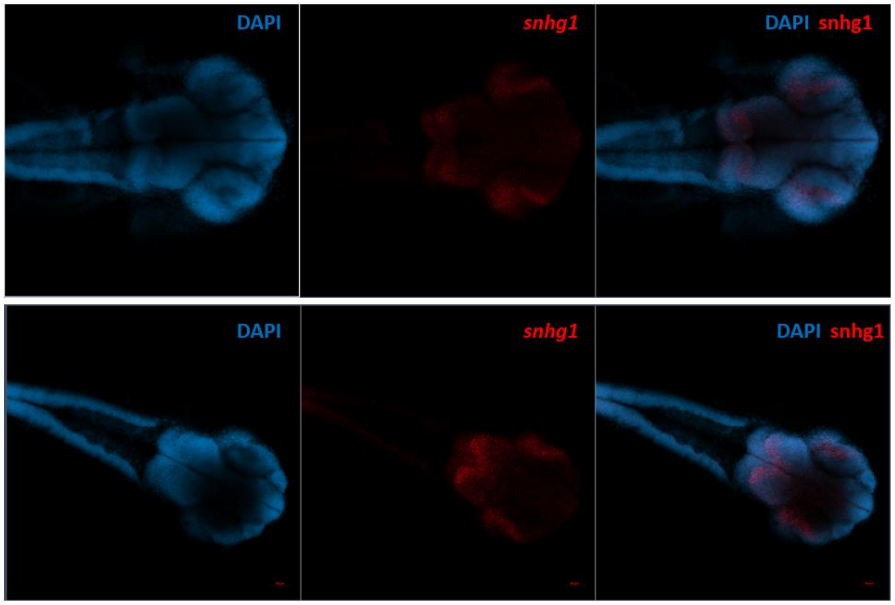


Figure 4.43. Brain-specific expression of *snhg1* gene. *A) Expression of snhg1 in comparison to dapB-negative control B) in zebrafish at 28 hours post fertilization. Fluorescent signal coming from RNAscope probes is shown in red, while DAPI signal in blue.*

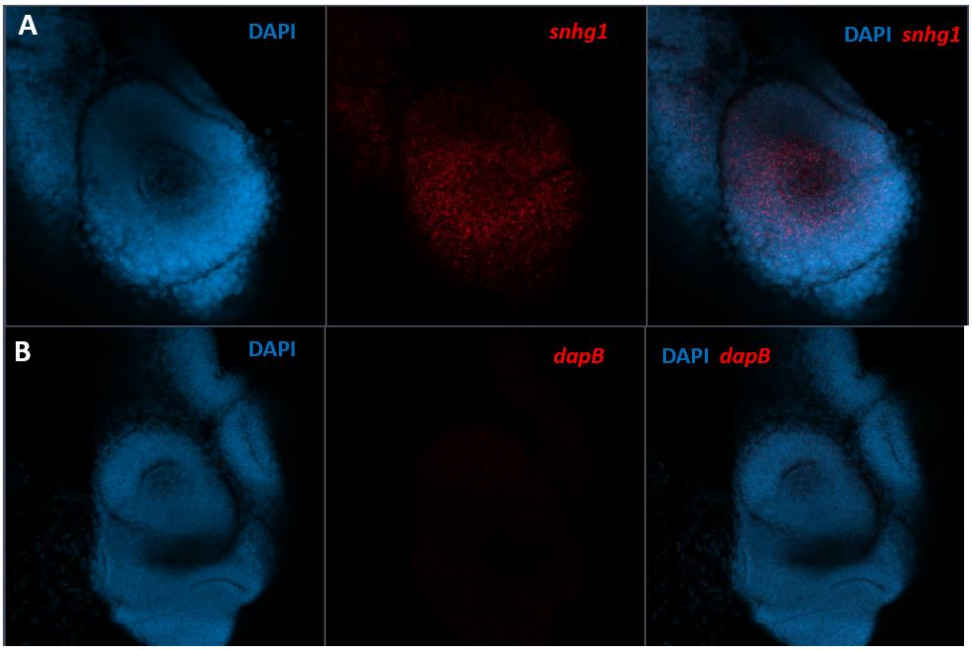


Figure 4.44. Eye-specific expression of *snhg1* gene. *A) Expression of snhg1 in comparison to dapB-negative control B) in zebrafish at 28 hours post fertilization. Fluorescent signal coming from RNAscope probes is shown in red, while DAPI signal in blue.*

This finding was further supported by data from Daniocell (Farrell et al., 2018; Sur et al., 2023) (Figure S12). Additionally, RNAscope results indicated that *snhg1* exhibited dual subcellular localization, with expression detected in both the cytoplasm and the nucleus (Figure 4.45).

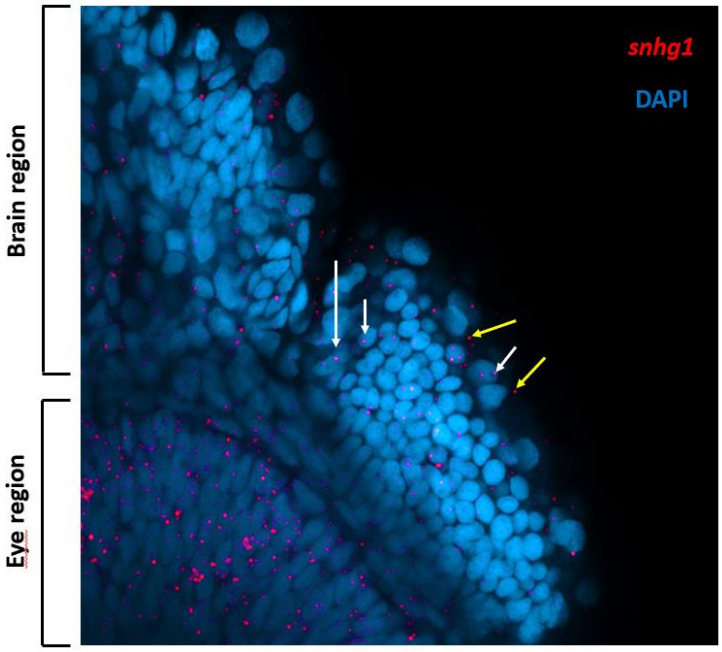


Figure 4.45. Subcellular localization of *snhg1* gene in zebrafish at 28 hours post fertilization. *Fluorescent signal coming from RNAscope probes is shown in red, while DAPI signal in blue. White arrows indicate nuclear RNA molecules, while yellow arrows denote cytoplasmic RNAs.*

5. DISCUSSION

The zebrafish is a powerful model for studying vertebrate biology and disease, providing significant opportunities to advance our understanding of lncRNA functionality. The functional characterization of lncRNAs is heavily dependent on the quality of genome annotations, highlighting the need for precise and comprehensive mapping of lncRNA genes. Unfortunately, current zebrafish lncRNA annotations are of lower quality compared to those of human and mouse genomes, with many incomplete or missing transcript models and unannotated lncRNA loci. The primary aim of this study was to establish a comprehensive and accurate lncRNA catalog for the zebrafish genome to enhance its value as a model for studying lncRNA functionality. To achieve this, the CapTrap-seq (Carbonell-Sala et al., 2024) full-length library preparation method was optimized for zebrafish to address the issue of incomplete transcript models. To evaluate its performance, CapTrap-seq was benchmarked against the widely used Template Switching (TSO) technology on a panel of zebrafish samples. Furthermore, the implementation of LyRic (Carbonell-Sala et al., 2024), a pipeline developed under the GENCODE consortium, enabled high-throughput and automated analysis of raw reads from long-read sequencing experiments, facilitating accurate and highly sensitive transcript model reconstruction with minimal human intervention.

The analysis demonstrated that CapTrap-seq outperforms TSO by producing higher-quality sequencing reads. The inferior quality of TSO libraries was evidenced by elevated sequencing error rates, particularly mismatches, a reduced proportion of High Confidence Genome Mappings (HCGMs), and a significantly higher fraction of less reliable monoexonic reads. This suggests that the TSO method can negatively affect annotation quality, as sequencing errors may lead to incorrect splice junctions and transcript artifacts (Zhang et al., 2022; Verwilt et al., 2023), while unvalidated unspliced reads can cause false-positive gene assignments, as shown by Zhang et al. (Zhang et al., 2022). Importantly, CapTrap-seq has demonstrated superior performance in enriching for full-length (5' and 3' complete) spliced molecules, reducing the issue of incomplete transcript termini and resulting in more accurate and complete transcript models. Furthermore, CapTrap-seq's ability to selectively target 5'-capped molecules resulted in a reduction of rRNA content to below 0.002%, ten times lower than TSO, even when combined with an additional rRNA removal step.

This advantage eliminates the need for costly, organism-specific rRNA removal processes, establishing CapTrap-seq as a versatile, cost-effective, and high-quality method for transcriptome analysis across virtually all eukaryotic organisms.

CapTrap-seq also has a few limitations. Its multistep procedure leads to the production of slightly shorter transcript models (100-300 bp) compared to TSO and also affects detection rates for spike-ins longer than 2 kb. Additionally, CapTrap-seq shows a lower proportion of polyadenylated reads than TSO, despite both methods using oligo-dT reverse transcription. Similar drawbacks have been reported in human samples (Carbonell-Sala et al., 2024).

Size selection is a widely used method to enhance long-read sequencing by enriching for longer cDNA molecules (Alfonso-Gonzalez et al., 2023). The application of bead-based size selection to remove cDNA fragments shorter than 500 bp (SS500) resulted in an increase in median transcript model length, extending it by approximately 100-200 bp compared to non-size-selected CapTrap-seq libraries. This approach also addressed another limitation of CapTrap-seq—the low proportion of polyadenylated reads—bringing it closer to the levels observed with the TSO method. Importantly, incorporation of the size-selection step did not compromise transcript model completeness. As a result, CapTrap-seq enabled more accurate annotation of transcription start sites (TSSs) and transcription termination sites (TTSs) for both protein-coding genes and lncRNAs, and facilitated the discovery of a higher number of novel spliced loci, particularly full-length ones, compared to TSO.

Although the results of CapTrap-seq optimization are highly promising, it is important to acknowledge the emergence of novel library preparation methods that offer comparable performance with less complexity and shorter protocols. For instance, a new Smartseq2 protocol modality introduced this year incorporates Terminator 5'-Phosphate-Dependent Exonuclease (TEX) enzyme treatment, achieving a similar proportion of 5' and 3' complete transcript models while yielding significantly longer transcript models compared to CapTrap-seq (Carbonell-Sala et al., 2024; Pardo-Palacios, et al., 2024). However, this protocol has so far only been tested on the PacBio platform, limiting a comprehensive assessment of its performance on the ONT platform. In comparison, the CapTrap-seq protocol has been effectively tested on both the PacBio and ONT platforms, demonstrating its versatility and reliability across different sequencing technologies.

The low expression of lncRNAs presents a significant challenge for accurate annotation, as it leads to their underrepresentation in sequencing libraries and resulting in gaps in reference annotations. Therefore to improve the annotation of lncRNAs in the zebrafish genome, I combined the CapTrap-seq SS500 library preparation method with the Capture Long-read Sequencing (CLS), a targeted RNA sequencing approach (Lagarde et al., 2017; Kaur et al., 2024). Moreover, to address the bias in zebrafish genome annotation toward developmental stages, I focused primarily on transcriptomically complex adult zebrafish tissues. The CapTrap-CLS approach in zebrafish outperformed both the original CLS protocol (Lagarde et al., 2017) and CapTrap-CLS (Kaur et al., 2024), achieving 18 to 100 times enrichment of targeted regions, depending on the tissue analyzed. This represents up to five-fold improvement over the CLS approach and about three times higher enrichment compared to CapTrap-CLS in human and mouse samples. This highlights that the CapTrap-CLS method can be easily optimized for use in other model organisms while maintaining high quality and efficiency.

The integration of full-length library preparation and long-read sequencing has substantially enhanced the zebrafish Ensembl annotation, uncovering over 14,000 novel genes and increasing the average number of isoforms per gene nearly fourfold. A protein-coding potential analysis determined that over 12,000 of these novel loci are predicted as lncRNAs, thereby expanding the existing zebrafish lncRNA catalog by four times. Furthermore, the application of CapTrap-CLS significantly contributed to the generation of high-quality, complete transcript models, with over 60% of models classified as complete, outpacing the GENCODE v47 lncRNA annotation for zebrafish. This comprehensive annotation provides a crucial resource for advancing zebrafish research, both for studying lncRNAs and other aspects of biological inquiry.

In this study, I produced the first set of full-length lncRNA models, providing a reliable foundation for testing the functional potential of lncRNA molecules. I aimed to assess how the application of CapTrap-CLS influenced the annotation of functional lncRNAs. To selectively identify these potentially functional lncRNAs, I employed a multistep selection criteria, starting with the ConnectOR synteny-based approach to identify positionally conserved lncRNAs between humans and zebrafish. This approach resulted in the identification of 49 zebrafish lncRNAs with corresponding human counterparts. Notably, several of these lncRNAs were small RNA host genes, including miRNA and snoRNA host genes, which are believed to be biologically relevant

and may play significant roles in gene regulation (Sun et al, 2021; Monziani and Ulitsky, 2023). I further investigated the association of these lncRNAs with human diseases and cancer and explored their expression patterns in zebrafish. The analysis revealed that these positionally conserved small RNA host genes were frequently associated with diseases and exhibited tissue- and development-specific expression patterns, particularly during brain and heart development, supporting their potential involvement in these processes. These examples highlight the important role of small RNA host genes in the development and function of various organs. However, it remains uncertain whether their function is limited to small RNAs or if the lncRNA host genes themselves may play a regulatory role in an RNA molecule-dependent manner. Notably, the extended annotation generated using CapTrap-CLS revealed a significant number of alternative processing events for these small non-coding RNAs, including the use of alternative transcription start sites (TSSs), transcription termination sites (TTSs), and alternative splicing mechanisms, such as exon skipping or intron retention, that affected the small RNA coding regions. Alternative processing is believed to play a crucial role in regulating expression levels, cellular fates and the functionality of both the host genes and the encoded small RNAs (Sun et al, 2021; Monziani and Ulitsky, 2023). However, the precise details of this regulatory process and its biological implications for lncRNA functionality remain to be elucidated. The extended Ensembl+ annotation offers significant potential to advance our understanding of these mechanisms.

To contribute to this investigation, I focused on the characterization of the enigmatic *snhg1* gene (snoRNA host gene 1), the highest-expressed positionally conserved lncRNA. First, I investigated its expression patterns of *snhg1* during organ development in zebrafish. RNAscope experiments at 1 day post-fertilization (dpf) revealed eye- and brain-specific expression patterns which were corroborated by data from Daniocell (Farrell et al., 2018; Sur et al., 2023). Additionally, *SNHG1* and its associated snoRNAs have been linked to ocular neovascularization and corneal angiogenesis (Liu et al., 2016; Hu et al. 2021). Given the conserved genomic position of *SNHG1* between humans and zebrafish, along with its eye-specific expression, zebrafish provide a valuable model for studying its regulatory role in eye development and eye-related diseases. RNAscope analysis also revealed that *snhg1* exhibits dual subcellular localization, being detected in both the cytoplasm and nucleus. However, the regulatory effects of intron retention or differential processing of *snhg1* and associated snoRNAs on their localization patterns remain to be fully understood.

In conclusion, the integration of the CapTrap-seq library preparation method with long-read RNA sequencing and LyRic analysis tools enabled the creation of highly accurate full-length transcript maps of the zebrafish genome, significantly reducing transcript termini incompleteness. Additionally, the combination of CapTrap-seq with the CLS approach resulted in a substantial improvement of lncRNA annotation for the zebrafish genome, leading to the identification of thousands of novel lncRNA loci and transcript isoforms. This extended lncRNA annotation will undoubtedly enhance our understanding of lncRNA functionality in vertebrates and provide a foundation for linking specific lncRNA processing events to their subsequent cellular fates.

6. CONCLUSIONS

1. The integration of the CapTrap-seq full-length library preparation method with long-read sequencing enables significant advancements in genome annotation facilitating the accurate mapping of transcription start and termination sites.
2. Incorporating a size selection step enhances CapTrap-seq performance in zebrafish by addressing its primary limitations. This includes enhancing the sequencing of longer cDNA molecules, increasing the proportion of polyadenylated reads, while simultaneously preserving the completeness of transcript models and improving the detection of novel spliced loci.
3. CapTrap-seq can be effectively combined with CLS-targeted RNA sequencing, offering a robust approach for improving lncRNA annotation in zebrafish.
4. The application of the CapTrap-CLS approach resulted in the creation of the first comprehensive full-length lncRNA dataset for the zebrafish genome. This dataset expanded lncRNA loci by fourfold and doubled the average number of transcript isoforms per gene compared to Ensembl v104.
5. Uniquely, the lncRNA dataset generated in this study specifically emphasizes positionally conserved genes, providing a valuable resource anticipated to significantly facilitate the functional characterization of lncRNAs across species.
6. Focusing on adult zebrafish tissues and organs reduced the bias in lncRNA annotation toward developmental samples, enabling more accurate comparisons between zebrafish and other organisms.
7. Finally, *snhg1*, characterized by its distinct expression patterns and positional conservation, has emerged as a promising target for functional analysis, with strong potential to be implicated in vertebrate eye development and eye-related diseases.

7. PERSPECTIVES

Long-read sequencing technologies have the potential to transform genomics and transcriptomics by overcoming the limitations of traditional short-read platforms, paving the way for major advancements in biological research. Although there has been ongoing debate about whether Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) will lead the sequencing field, growing evidence indicates that these platforms serve complementary rather than competing roles. Utilizing both systems in combination can deliver the most precise and comprehensive insights (Nurk et al., 2022; Carbonell-Sala et al., 2024; Pardo-Palacios et al., 2024), particularly for complex genomes and transcriptomes, as each offers unique strengths in read length, accuracy, and cost. This synergistic approach is expected to define the future of long-read sequencing.

Long-read sequencing is starting to reveal its transformative potential in zebrafish research, though its use remains limited. Beyond its role in transcriptomic studies, long-read sequencing holds exceptional promise for completing zebrafish genome assemblies. Unlike the human genome, the zebrafish genome contains hundreds of unresolved gaps (Chernyavskaya et al., 2022), potentially containing unannotated genes, particularly non-coding ones. These genes are often located in complex, repetitive regions that short-read sequencing has difficulty resolving. As seen in the success of telomere-to-telomere (T2T) sequencing in humans, where nearly 2,000 genomic regions corresponding to previously uncharacterized genes were identified, with the majority being non-coding (Nurk et al., 2022), long-read sequencing could similarly drive breakthroughs in zebrafish genomics. Improved accessibility and optimization of these technologies may also enable the creation of a zebrafish pangenome, capturing strain-specific genetic diversity and providing a richer reference for the research community, especially that zebrafish T2T assembly is coming.

Lower sequencing costs, improved technologies, and the creation of user-friendly, automated bioinformatics tools are essential for fully realizing the potential of long-read sequencing in zebrafish and beyond. Broader adoption of these technologies will advance functional genomics, enhance disease modeling, and deepen our understanding of evolutionary biology. These combined advancements will enable groundbreaking discoveries, not only in zebrafish but across a wide range of species, offering unprecedented insights into complex genomes and transcriptomes.

8. DATA AVAILABILITY

All relevant information for this study, including the PDF version of the thesis, figures, input data, and code used for analysis, is available at https://github.com/cobRNA/MKwiatkowska_PhD_Thesis.

BIBLIOGRAPHY

Alfonso-Gonzalez C, Arrigoni L, Ozbulut HC, Falk S, Bönisch U, Hilgers V. Identification of regulatory links between transcription and RNA processing with long-read sequencing. *STAR Protoc*. 2023;4(4):102505. doi:10.1016/j.xpro.2023.102505

Ali T, Grote P. Beyond the RNA-dependent function of LncRNA genes. *Elife*. 2020;9:e60583. Published 2020 Oct 23. doi:10.7554/eLife.60583

Amaral P, Carbonell-Sala S, De La Vega FM, et al. The status of the human gene catalogue. *Nature*. 2023;622(7981):41-47. doi:10.1038/s41586-023-06490-x

Amin, N., McGrath, A. & Chen, YP.P. Evaluation of deep learning in non-coding RNA classification. *Nat Mach Intell.* 2019;1:246–256. https://doi.org/10.1038/s42256-019-0051-2

Andergassen D, Rinn JL. From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo. *Nat Rev Genet*. 2022;23(4):229-243. doi:10.1038/s41576-021-00427-8

Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, Dong D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res*. 2019;47(D1):D1034-D1037. doi:10.1093/nar/gky905

Basu K, Dey A, Kiran M. Inefficient splicing of long non-coding RNAs is associated with higher transcript complexity in human and mouse. *RNA Biol*. 2023;20(1):563-572. doi:10.1080/15476286.2023.2242649

Brachet J. La localisation des acides pentose nucléiques dans les tissus animaux et les oeufs d'Amphibiens en voie de dévelopepment. Arch. Biol. 1942; 53:207-257.

Bridges MC, Daulagala AC, Kourtidis A. LNCcation: lncRNA localization and function. *J Cell Biol*. 2021;220(2):e202009045. doi:10.1083/jcb.202009045

Bryzghalov O, Makałowska I, Szcześniak MW. lncEvo: automated identification and conservation study of long noncoding RNAs. *BMC Bioinformatics*. 2021;22(1):59. Published 2021 Feb 9. doi:10.1186/s12859-021-03991-2

Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;25(18):1915-1927. doi:10.1101/gad.17446611

Carbonell-Sala S, Guigó R. 5' Capping protocol to add 5’ cap structures to exogenous synthetic RNA references (spike-ins), 24 May 2024, PROTOCOL (Version 1) available at Protocol Exchange doi:https://doi.org/10.21203/rs.3.pex-2649/v1

Carbonell-Sala S, Guigó R. CapTrap-Seq cDNA library preparation for full-length RNA sequencing, 24 May 2024, PROTOCOL (Version 1) available at Protocol Exchange doi:https://doi.org/10.21203/rs.3.pex-2646/v1

Carbonell-Sala S, Lagarde J, Nishiyori H, et al. CapTrap-Seq: A platform-agnostic and quantitative approach for high-fidelity full-length RNA transcript sequencing. Preprint. *bioRxiv*. 2023;2023.06.16.543444. doi:10.1101/2023.06.16.543444

Carbonell-Sala S, Perteghella T, Lagarde J, et al. CapTrap-seq: a platform-agnostic and quantitative approach for high-fidelity full-length RNA sequencing. *Nat Commun*. 2024;15(1):5278. Published 2024 Jun 27. doi:10.1038/s41467-024-49523-3

Carbonell Sala S, Uszczyńska-Ratajczak B, Lagarde J, Johnson R, Guigó R. Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing (CLS). *Methods Mol Biol*. 2021;2254:133-159. doi:10.1007/978-1-0716-1158-6_9

Carlevaro-Fita J, Johnson R. Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization. *Mol Cell*. 2019;73(5):869-883. doi:10.1016/j.molcel.2019.02.008

Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res*. 2019;29(2):208-222. doi:10.1101/gr.229922.117

Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. *Science*. 2005;309(5740):1559-1563. doi:10.1126/science.1112014

Caspersson T. The relations between nucleic acid and protein synthesis. *Symp Soc Exp Biol*. 1947;(1):127-151.

Chen H, Du G, Song X, Li L. Non-coding Transcripts from Enhancers: New Insights into Enhancer Activity and Gene Expression Regulation. *Genomics Proteomics Bioinformatics*. 2017;15(3):201-207. doi:10.1016/j.gpb.2017.02.003

Chernyavskaya Y, Zhang X, Liu J, Blackburn J. Long-read sequencing of the zebrafish genome reorganizes genomic architecture. *BMC Genomics*. 2022;23(1):116. Published 2022 Feb 10. doi:10.1186/s12864-022-08349-3

Chillón I, Marcia M. The molecular structure of long non-coding RNAs: emerging patterns and functional implications. *Crit Rev Biochem Mol Biol*. 2020;55(6):662-690. doi:10.1080/10409238.2020.1828259

Choi TY, Choi TI, Lee YR, Choe SK, Kim CH. Zebrafish as an animal model for biomedical research. *Exp Mol Med*. 2021;53(3):310-317. doi:10.1038/s12276-021-00571-5

Cobb M. Who discovered messenger RNA?. *Curr Biol*. 2015;25(13):R526-R532. doi:10.1016/j.cub.2015.05.032

Cobb M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol*. 2017;15(9):e2003243. Published 2017 Sep 18. doi:10.1371/journal.pbio.2003243

Constanty F, Shkumatava A. lncRNAs in development and differentiation: from sequence motifs to functional characterization. *Development*. 2021;148(1):dev182741. Published 2021 Jan 13. doi:10.1242/dev.182741

Crick, F. (1958). On Protein Synthesis. In Symposium of the Society for Experimental Biology (Vol. 12, pp. 139-163). Cambridge University Press
And 20 years later: In PLos Maddox, B. (2017).
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2003243

Crick F. Central dogma of molecular biology. *Nature*. 1970;227(5258):561-563. doi:10.1038/227561a0 ‘

Das T, Deb A, Parida S, Mondal S, Khatua S, Ghosh Z. LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients. *RNA Biol*. 2021;18(8):1136-1151. doi:10.1080/15476286.2020.1833529

De Paoli-Iseppi R, Gleeson J, Clark MB. Isoform Age - Splice Isoform Profiling Using Long-Read Technologies. *Front Mol Biosci*. 2021;8:711733. doi:10.3389/fmolb.2021.711733

Deveson IW, Brunck ME, Blackburn J, et al. Universal Alternative Splicing of Noncoding Exons. *Cell Syst*. 2018;6(2):245-255.e5. doi:10.1016/j.cels.2017.12.005

Diederichs S. The four dimensions of noncoding RNA conservation. *Trends Genet*. 2014;30(4):121-123. doi:10.1016/j.tig.2014.01.004

DiStefano JK. The Emerging Role of Long Noncoding RNAs in Human Disease. *Methods Mol Biol*. 2018;1706:91-110. doi:10.1007/978-1-4939-7471-9_6

Dhiman H, Kapoor S, Sivadas A, Sivasubbu S, Scaria V. zflncRNApedia: A Comprehensive Online Resource for Zebrafish Long Non-Coding RNAs. *PLoS One*. 2015;10(6):e0129997. doi:10.1371/journal.pone.0129997

Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature*. 2012;489(7414):101-108. doi:10.1038/nature11233

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

Esposito R, Lanzós A, Uroda T, et al. Tumour mutations in long noncoding RNAs enhance cell fitness. *Nat Commun*. 2023;14(1):3342. Published 2023 Jun 8. doi:10.1038/s41467-023-39160-7

Ezkurdia I, Juan D, Rodriguez JM, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet*. 2014;23(22):5866-5878. doi:10.1093/hmg/ddu309

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*. 2018;360(6392):eaar3131. doi:10.1126/science.aar3131

Fort V, Khelifi G, Hussein SMI. Long non-coding RNAs and transposable elements: A functional relationship. *Biochim Biophys Acta Mol Cell Res*. 2021;1868(1):118837. doi:10.1016/j.bbamcr.2020.118837

Franke A, Baker BS. The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. *Mol Cell*. 1999;4(1):117-122. doi:10.1016/s1097-2765(00)80193-8

Frankish A, Carbonell-Sala S, Diekhans M, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res*. 2023;51(D1):D942-D949. doi:10.1093/nar/gkac1071

Gabory A, Jammes H, Dandolo L. The H19 locus: role of an imprinted non-coding RNA in growth and development. *Bioessays*. 2010;32(6):473-480. doi:10.1002/bies.200900170

Grant J, Mahadevaiah SK, Khil P, et al. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature*. 2012;487(7406):254-258. doi:10.1038/nature11171

Grünberger F, Ferreira-Cerca S, Grohmann D. Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*. *RNA*. 2022;28(3):400-417. doi:10.1261/rna.078937.121

Guo CJ, Xu G, Chen LL. Mechanisms of Long Noncoding RNA Nuclear Retention. *Trends Biochem Sci*. 2020;45(11):947-960. doi:10.1016/j.tibs.2020.07.001

Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature*. 2012;482(7385):339-346. Published 2012 Feb 15. doi:10.1038/nature10887

Harris KA, Breaker RR. Large Noncoding RNAs in Bacteria. *Microbiol Spectr*. 2018;6(4):10.1128/microbiolspec.RWR-0005-2017 doi:10.1128/microbiolspec.RWR-0005-2017

Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucleic Acids Res*. 2024;52(D1):D891-D899. doi:10.1093/nar/gkad1049

Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol*. 2006;7 Suppl 1(Suppl 1):S4.1-S4.9. doi:10.1186/gb-2006-7-s1-s4

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*. 2015;11(7):1110-1122. doi:10.1016/j.celrep.2015.04.023

Hombach S, Kretz M. Non-coding RNAs: Classification, Biology and Functioning. *Adv Exp Med Biol*. 2016;937:3-17. doi:10.1007/978-3-319-42059-2_1

Hong YK, Ontiveros SD, Strauss WM. A revision of the human XIST gene organization and structural comparison with mouse Xist. Mamm Genome. 2000;11(3):220-224. doi:10.1007/s003350010040

Howe K, Clark MD, Torroja CF, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 2013;496(7446):498-503. doi:10.1038/nature12111

Hu X, Xing H, Wang X, et al. Knockdown of LncRNA SNHG1 Suppresses Corneal Angiogenesis by the Regulation of miR-195-5p/VEGF-A. *J Ophthalmol*. 2021;2021:6646512. Published 2021 Oct 19. doi:10.1155/2021/6646512

Hu X, Chen W, Li J, et al. ZFLNC: a comprehensive and well-annotated database for zebrafish lncRNA. *Database (Oxford)*. 2018;2018:bay114. Published 2018 Jan 1. doi:10.1093/database/bay114

Huang W, Xiong T, Zhao Y, et al. Computational prediction and experimental validation identify functionally conserved lncRNAs from zebrafish to human. *Nat Genet*. 2024;56(1):124-135. doi:10.1038/s41588-023-01620-7

Huarte M. The emerging role of lncRNAs in cancer. *Nat Med*. 2015;21(11):1253-1261. doi:10.1038/nm.3981

Ibrahim F, Oppelt J, Maragkakis M, Mourelatos Z. TERA-Seq: true end-to-end sequencing of native RNA molecules for transcriptome characterization. *Nucleic Acids Res*. 2021;49(20):e115. doi:10.1093/nar/gkab713

Iyer MK, Niknafs YS, Malik R, et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*. 2015;47(3):199-208. doi:10.1038/ng.3192

Jandura A, Krause HM. The New RNA World: Growing Evidence for Long Noncoding RNA Functionality. *Trends Genet*. 2017;33(10):665-676. doi:10.1016/j.tig.2017.08.002

Jarroux J, Morillon A, Pinskaya M. History, Discovery, and Classification of lncRNAs. *Adv Exp Med Biol*. 2017;1008:1-46. doi:10.1007/978-981-10-5203-3_1

Jiao F, Pahwa K, Manning M, Dochy N, Geuten K. Cold Induced Antisense Transcription of *FLOWERING LOCUS C* in Distant Grasses. *Front Plant Sci*. 2019;10:72. Published 2019 Feb 1. doi:10.3389/fpls.2019.00072

Johnson R, Guigó R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*. 2014;20(7):959-976. doi:10.1261/rna.044560.114

Kang YJ, Yang DC, Kong L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*. 2017;45(W1):W12-W16. doi:10.1093/nar/gkx428

Karner H, Webb CH, Carmona S, et al. Functional Conservation of LncRNA JPX Despite Sequence and Structural Divergence. *J Mol Biol*. 2020;432(2):283-300. doi:10.1016/j.jmb.2019.09.002

Kaur G, Perteghella T, Carbonell-Sala S, et al. GENCODE: massively expanding the lncRNA catalog through capture long-read RNA sequencing. Preprint. *bioRxiv*. 2024;2024.10.29.620654. Published 2024 Oct 31. doi:10.1101/2024.10.29.620654

Kaushik K, Leonard VE, Kv S, et al. Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PLoS One*. 2013;8(12):e83616. Published 2013 Dec 31. doi:10.1371/journal.pone.0083616

Khan M, Hou S, Chen M, Lei H. Mechanisms of RNA export and nuclear retention. *Wiley Interdiscip Rev RNA*. 2023;14(3):e1755. doi:10.1002/wrna.1755

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. *Dev Dyn*. 1995;203(3):253-310. doi:10.1002/aja.1002030302

Kirk JM, Kim SO, Inoue K, et al. Functional classification of long non-coding RNAs by k-mer content. *Nat Genet*. 2018;50(10):1474-1482. doi:10.1038/s41588-018-0207-8

Kopp F, Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*. 2018;172(3):393-407. doi:10.1016/j.cell.2018.01.011

Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol*. 2013;11:59. Published 2013 May 30. doi:10.1186/1741-7007-11-59

Kutter C, Watt S, Stefflova K, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet*. 2012;8(7):e1002841. doi:10.1371/journal.pgen.1002841

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet*. 2017;49(12):1731-1740. doi:10.1038/ng.3988

Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062

Lawson ND, Li R, Shin M, et al. An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *Elife*. 2020;9:e55792. Published 2020 Aug 24. doi:10.7554/eLife.55792

Lee S, Kopp F, Chang TC, et al. Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell*. 2016;164(1-2):69-80. doi:10.1016/j.cell.2015.12.017

Lee H, Zhang Z, Krause HM. Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners?. *Trends Genet*. 2019;35(12):892-902. doi:10.1016/j.tig.2019.09.006

Lewin HA, Richards S, Lieberman Aiden E, et al. The Earth BioGenome Project 2020: Starting the clock. *Proc Natl Acad Sci U S A*. 2022;119(4):e2115635118. doi:10.1073/pnas.2115635118

Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A*. 2018;115(17):4325-4333. doi:10.1073/pnas.1720115115

Li Z, Liu L, Feng C, et al. LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations. *Nucleic Acids Res*. 2023;51(D1):D186-D191. doi:10.1093/nar/gkac999

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet*. 2000;25(2):239-240. doi:10.1038/76126

Lieschke GJ, Currie PD. Animal models of human disease: zebrafish swim into view. *Nat Rev Genet*. 2007;8(5):353-367. doi:10.1038/nrg2091

Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27(13):i275-i282. doi:10.1093/bioinformatics/btr209

Liu CH, Wang Z, Sun Y, SanGiovanni JP, Chen J. Retinal expression of small non-coding RNAs in a murine model of proliferative retinopathy. *Sci Rep*. 2016;6:33947. Published 2016 Sep 22. doi:10.1038/srep33947

Lorenzi L, Chiu HS, Avila Cobos F, et al. The RNA Atlas expands the catalog of human non-coding RNAs. *Nat Biotechnol*. 2021;39(11):1453-1465. doi:10.1038/s41587-021-00936-1

Loveland J. VEGA, the genome browser with a difference. *Brief Bioinform*. 2005;6(2):189-193. doi:10.1093/bib/6.2.189

Lyu QR, Zhang S, Zhang Z, Tang Z. Functional knockout of long non-coding RNAs with genome editing. *Front Genet*. 2023;14:1242129. Published 2023 Aug 29. doi:10.3389/fgene.2023.1242129

Ma L, Cao J, Liu L, et al. LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res*. 2019;47(5):2699. doi:10.1093/nar/gkz073

Machikhin AS, Volkov MV, Burlakov AB, Khokhlov DD, Potemkin AV. Blood Vessel Imaging at Pre-Larval Stages of Zebrafish Embryonic Development. *Diagnostics (Basel)*. 2020;10(11):886. Published 2020 Oct 30. doi:10.3390/diagnostics10110886

Marchese FP, Raimondi I, Huarte M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol*. 2017;18(1):206. Published 2017 Oct 31. doi:10.1186/s13059-017-1348-2

Mattick JS. The central role of RNA in human development and cognition. *FEBS Lett*. 2011;585(11):1600-1616. doi:10.1016/j.febslet.2011.05.001

Mattick J, Amaral P. RNA, the Epicenter of Genetic Information: A new understanding of molecular biology. Abingdon (UK): CRC Press; 2022 Sep 20. Chapter 13, Large RNAs with Many Functions. Available from: https://www.ncbi.nlm.nih.gov/books/NBK595947/ doi: 10.1201/9781003109242-13

Mattick JS, Amaral PP, Carninci P, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol*. 2023;24(6):430-447. doi:10.1038/s41580-022-00566-8

McDonnell E, Strasser K, Tsang A. Manual Gene Curation and Functional Annotation. *Methods Mol Biol*. 2018;1775:185-208. doi:10.1007/978-1-4939-7804-5_16

Mehjabin R, Xiong L, Huang R, et al. Full-Length Transcriptome Sequencing and the Discovery of New Transcripts in the Unfertilized Eggs of Zebrafish (*Danio rerio*). *G3 (Bethesda)*. 2019;9(6):1831-1838. Published 2019 Jun 5. doi:10.1534/g3.119.200997

Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet*. 2009;10(3):155-159. doi:10.1038/nrg2521

Monziani A, Ulitsky I. Noncoding snoRNA host genes are a distinct subclass of long noncoding RNAs. *Trends Genet*. 2023;39(12):908-923. doi:10.1016/j.tig.2023.09.001

Moore JB 4th, Uchida S. Functional characterization of long noncoding RNAs. *Curr Opin Cardiol*. 2020;35(3):199-206. doi:10.1097/HCO.0000000000000725

Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. 2022;604(7905):310-315. doi:10.1038/s41586-022-04558-8

Noh JH, Kim KM, McClusky WG, Abdelmohsen K, Gorospe M. Cytoplasmic functions of long noncoding RNAs. *Wiley Interdiscip Rev RNA*. 2018;9(3):e1471. doi:10.1002/wrna.1471

Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput Struct Biotechnol J*. 2023;21:2352-2364. Published 2023 Mar 24. doi:10.1016/j.csbj.2023.03.038

Nowacka M, Latoch P, Izert MA, et al. A cap 0-dependent mRNA capture method to analyze the yeast transcriptome. *Nucleic Acids Res*. 2022;50(22):e132. doi:10.1093/nar/gkac903

Nudelman G, Frasca A, Kent B, et al. High resolution annotation of zebrafish transcriptome using long-read sequencing. *Genome Res*. 2018;28(9):1415-1425. doi:10.1101/gr.223586.117

Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44-53. doi:10.1126/science.abj6987

Ohno S. So much "junk" DNA in our genome. *Brookhaven Symp Biol*. 1972;23:366-370.

O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189

Ounzain S, Micheletti R, Arnan C, et al. CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis. *J Mol Cell Cardiol*. 2015;89(Pt A):98-112. doi:10.1016/j.yjmcc.2015.09.016

Page ML, Aguzzoli Heberle B, Brandon JA, et al. Surveying the landscape of RNA isoform diversity and expression across 9 GTEx tissues using long-read sequencing data. Preprint. *bioRxiv*. 2024;2024.02.13.579945. doi:10.1101/2024.02.13.579945

Palazzo AF, Koonin EV. Functional Long Non-coding RNAs Evolve from Junk Transcripts. *Cell*. 2020;183(5):1151-1161. doi:10.1016/j.cell.2020.09.047

Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk?. *Front Genet*. 2015;6:2. Published 2015 Jan 26. doi:10.3389/fgene.2015.00002

Pan J, Wang R, Shang F, Ma R, Rong Y, Zhang Y. Functional Micropeptides Encoded by Long Non-Coding RNAs: A Comprehensive Review. *Front Mol Biosci*. 2022;9:817517. Published 2022 Jun 13. doi:10.3389/fmolb.2022.817517

Pardo-Palacios FJ, Wang D, Reese F, *et al*. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* 2024;21:1349–1363. https://doi.org/10.1038/s41592-024-02298-3

Pauli A, Valen E, Lin MF, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*. 2012;22(3):577-591. doi:10.1101/gr.133009.111

Perron U, Provero P, Molineris I. In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression. *BMC Bioinformatics*. 2017;18(Suppl 5):144. Published 2017 Mar 23. doi:10.1186/s12859-017-1535-x

Peukert D, Weber S, Lumsden A, Scholpp S. Lhx2 and Lhx9 determine neuronal differentiation and compartition in the caudal forebrain by regulating Wnt signaling. *PLoS Biol*. 2011;9(12):e1001218. doi:10.1371/journal.pbio.1001218

Pi-Roig A, Martin-Blanco E, Minguillon C. Distinct tissue-specific requirements for the zebrafish tbx5 genes during heart, retina and pectoral fin development. *Open Biol*. 2014;4(4):140014. Published 2014 Apr 23. doi:10.1098/rsob.140014

Plaisance I, Chouvardas P, Sun Y, et al. A transposable element into the human long noncoding RNA CARMEN is a switch for cardiac precursor cell specification. *Cardiovasc Res*. 2023;119(6):1361-1376. doi:10.1093/cvr/cvac191

Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell*. 2009;136(4):629-641. doi:10.1016/j.cell.2009.02.006

Rahman Khan, F., & Sulaiman Alhewairini, S. Zebrafish (Danio rerio) as a Model Organism. IntechOpen. 2019. doi: 10.5772/intechopen.81517

Ross CJ, Rom A, Spinrad A, Gelbard-Solodkin D, Degani N, Ulitsky I. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biol*. 2021;22(1):29. Published 2021 Jan 11. doi:10.1186/s13059-020-02247-1

Sahakyan A, Yang Y, Plath K. The Role of Xist in X-Chromosome Dosage Compensation. *Trends Cell Biol*. 2018;28(12):999-1013. doi:10.1016/j.tcb.2018.05.005

Shehwana H, Konu O. Comparative Transcriptomics Between Zebrafish and Mammals: A Roadmap for Discovery of Conserved and Unique Signaling Pathways in Physiology and Disease. *Front Cell Dev Biol*. 2019;7:5. doi:10.3389/fcell.2019.00005

Siniscalchi C, Di Palo A, Russo A, Potenza N. The lncRNAs at X Chromosome Inactivation Center: Not Just a Matter of Sex Dosage Compensation. *Int J Mol Sci*. 2022;23(2):611. Published 2022 Jan 6. doi:10.3390/ijms23020611

Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*. 2021;22(2):96-118. doi:10.1038/s41580-020-00315-9

Sun Q, Song YJ, Prasanth KV. One locus with two roles: microRNA-independent functions of microRNA-host-gene locus-encoded long noncoding RNAs. *Wiley Interdiscip Rev RNA*. 2021;12(3):e1625. doi:10.1002/wrna.1625

Sur A, Wang Y, Capar P, Margolin G, Prochaska MK, Farrell JA. Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. *Dev Cell*. 2023;58(24):3028-3047.e12. doi:10.1016/j.devcel.2023.11.001

Szcześniak MW, Wanowska E, Mukherjee N, Ohler U, Makałowska I. Towards a deeper annotation of human lncRNAs. *Biochim Biophys Acta Gene Regul Mech*. 2020;1863(4):194385. doi:10.1016/j.bbagrm.2019.05.003

Szcześniak MW, Kubiak MR, Wanowska E, Makałowska I. Comparative genomics in the search for conserved long noncoding RNAs. *Essays Biochem*. 2021;65(4):741-749. doi:10.1042/EBC20200069

Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet*. 2016;17(10):601-614. doi:10.1038/nrg.2016.85

Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell*. 2013;154(1):26-46. doi:10.1016/j.cell.2013.06.020

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*. 2011;147(7):1537-1550. doi:10.1016/j.cell.2011.11.055

Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet*. 2018;19(9):535-548. doi:10.1038/s41576-018-0017-y

Vancura A, Gutierrez AH, Hennig T, et al. Is Evolutionary Conservation a Useful Predictor for Cancer Long Noncoding RNAs? Insights from the Cancer LncRNA Census 3. *Noncoding RNA*. 2022;8(6):82. Published 2022 Dec 7. doi:10.3390/ncrna8060082

Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304-1351. doi:10.1126/science.1058040

Verwilt J, Mestdagh P, Vandesompele J. Artifacts and biases of the reverse transcription reaction in RNA sequencing. *RNA*. 2023;29(7):889-897. doi:10.1261/rna.079623.123

Vesterlund L, Jiao H, Unneberg P, Hovatta O, Kere J. The zebrafish transcriptome during early development. *BMC Dev Biol*. 2011;11:30. Published 2011 May 24. doi:10.1186/1471-213X-11-30

Volkin E, Astrachan L. Phosphorus incorporation in Escherichia coli ribo-nucleic acid after infection with bacteriophage T2. *Virology*. 1956;2(2):149-161. doi:10.1016/0042-6822(56)90016-2

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41(6):e74. doi:10.1093/nar/gkt006

Wang Z, Zhao Y, Zhang Y. Viral lncRNA: A regulatory molecule for controlling virus life cycle. *Noncoding RNA Res*. 2017;2(1):38-44. Published 2017 Mar 23. doi:10.1016/j.ncrna.2017.03.002

Weghorst F, Torres Marcén M, Faridi G, Lee YCG, Cramer KS. Deep Conservation and Unexpected Evolutionary History of Neighboring lncRNAs MALAT1 and NEAT1. *J Mol Evol*. 2024;92(1):30-41. doi:10.1007/s00239-023-10151-y

Werner S, Schmidt L, Marchand V, et al. Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. *Nucleic Acids Res*. 2020;48(7):3734-3746. doi:10.1093/nar/gkaa113

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*. 2008;36(Database issue):D753-D760. doi:10.1093/nar/gkm987

White RJ, Collins JE, Sealy IM, et al. A high-resolution mRNA expression time course of embryonic development in zebrafish. *Elife*. 2017;6:e30860. Published 2017 Nov 16. doi:10.7554/eLife.30860

Wohlers I, Garg S, Hehir-Kwa JY. Editorial: Long-read sequencing-Pitfalls, benefits and success stories. *Front Genet*. 2023;13:1114542. Published 2023 Jan 4. doi:10.3389/fgene.2022.1114542

Xavier Sabaté-Cadenas, Perrine Lavalou, Caroline Jane Ross, Lee Chen, Dina Zielinski, Sophie Vacher, Mireille Ledevin, Thibaut Larcher, Matthieu Petitjean, Louise Damy, Nicolas Servant, Ivan Bièche, Igor Ulitsky, Alena Shkumatava. Conserved RNA-binding protein interactions mediate syntologous lncRNA functions. bioRxiv 2024.08.21.605776; doi: https://doi.org/10.1101/2024.08.21.605776

Xiong P, Schneider RF, Hulsey CD, Meyer A, Franchini P. Conservation and novelty in the microRNA genomic landscape of hyperdiverse cichlid fishes. *Sci Rep*. 2019;9(1):13848. Published 2019 Sep 25. doi:10.1038/s41598-019-50124-0

Xiao Y, Ren Y, Hu W, et al. Long non-coding RNA-encoded micropeptides: functions, mechanisms and implications. Cell Death Discov. 2024;10(1):450. Published 2024 Oct 23. doi:10.1038/s41420-024-02175-0

Yadav VK, Jalmi SK, Tiwari S, Kerkar S. Deciphering shared attributes of plant long non-coding RNAs through a comparative computational approach. *Sci Rep*. 2023;13(1):15101. Published 2023 Sep 12. doi:10.1038/s41598-023-42420-7

Yang Y, Wang D, Miao YR, et al. lncRNASNP v3: an updated database for functional variants in long non-coding RNAs. *Nucleic Acids Res*. 2023;51(D1):D192-D198. doi:10.1093/nar/gkac981

Zhang R, Kuo R, Coulter M, et al. A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. *Genome Biol*. 2022;23(1):149. doi:10.1186/s13059-022-02711-0

Zhang X, Wang W, Zhu W, et al. Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *Int J Mol Sci*. 2019;20(22):5573. doi:10.3390/ijms20225573

Zhang Y, He Q, Hu Z, et al. Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat Struct Mol Biol*. 2016;23(6):522-530. doi:10.1038/nsmb.3211

Zhao Z, Zhang D, Yang F, et al. Evolutionarily conservative and non-conservative regulatory networks during primate interneuron development revealed by single-cell RNA and ATAC sequencing. *Cell Res*. 2022;32(5):425-436. doi:10.1038/s41422-022-00635-9

Zhou B, Ji B, Liu K, et al. EVLncRNAs 2.0: an updated database of manually curated functional long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res*. 2021;49(D1):D86-D91. doi:10.1093/nar/gkaa1076

Zhou C, Li M, Sun Y, Sultan Y, Li X. Systematic Identification of Long Noncoding RNAs during Three Key Organogenesis Stages in Zebrafish. *Int J Mol Sci*. 2024;25(6):3440. Published 2024 Mar 19. doi:10.3390/ijms25063440

Zhao L, Wang J, Li Y, et al. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res*. 2021;49(D1):D165-D171. doi:10.1093/nar/gkaa1046

Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res*. 2012;40(7):e54. doi:10.1093/nar/gkr1263

SUPPLEMENTARY MATERIALS

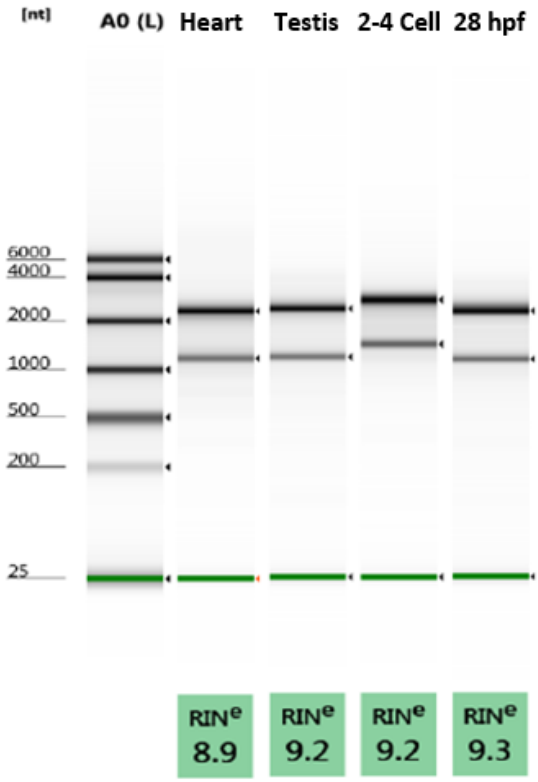


Figure S1. Integrity of RNA isolated from adult organs and zebrafish developmental stages.

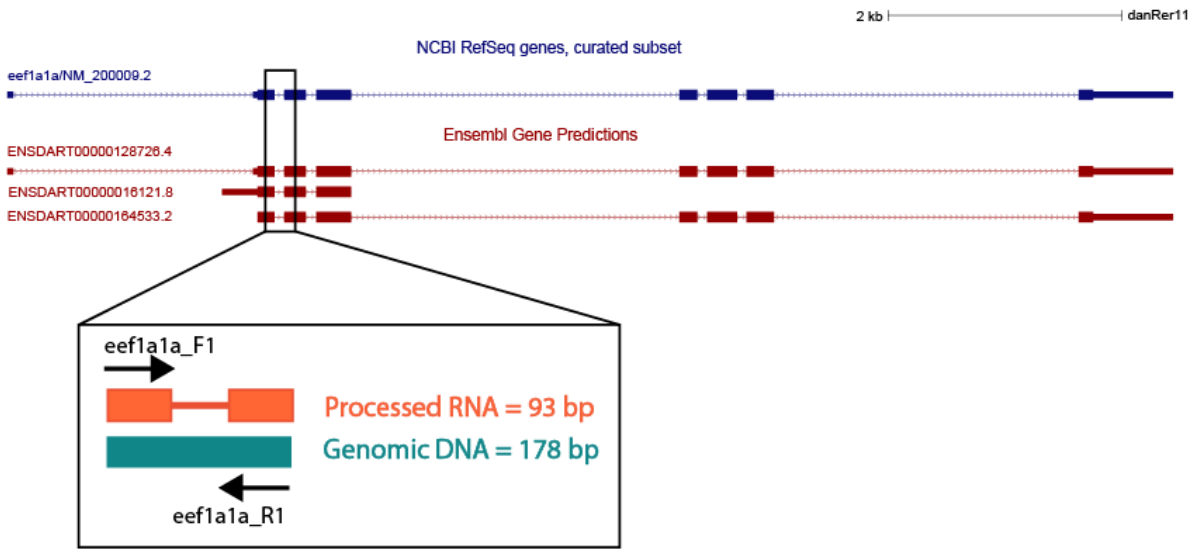


Figure S2. Graphical representation of the experiment designed to assess genomic DNA contamination.

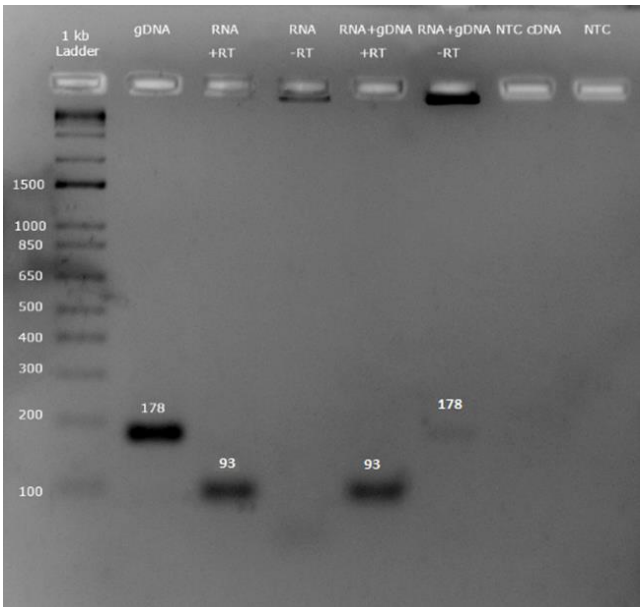


Figure S3. Performance of the RT-PCR method in detecting residual genomic DNA contamination. *gDNA indicates the sample where genomic DNA was used as a template for the PCR reaction; +RT represents the sample with Reverse Transcriptase included; -RT indicates the sample without Reverse Transcriptase; RNA + gDNA refers to the sample where RNA was artificially spiked with genomic DNA; NTC cDNA denotes the negative control in which water was added instead of RNA during the reverse transcription reaction; and NTC indicates the negative control for PCR, where water was added instead of cDNA.*

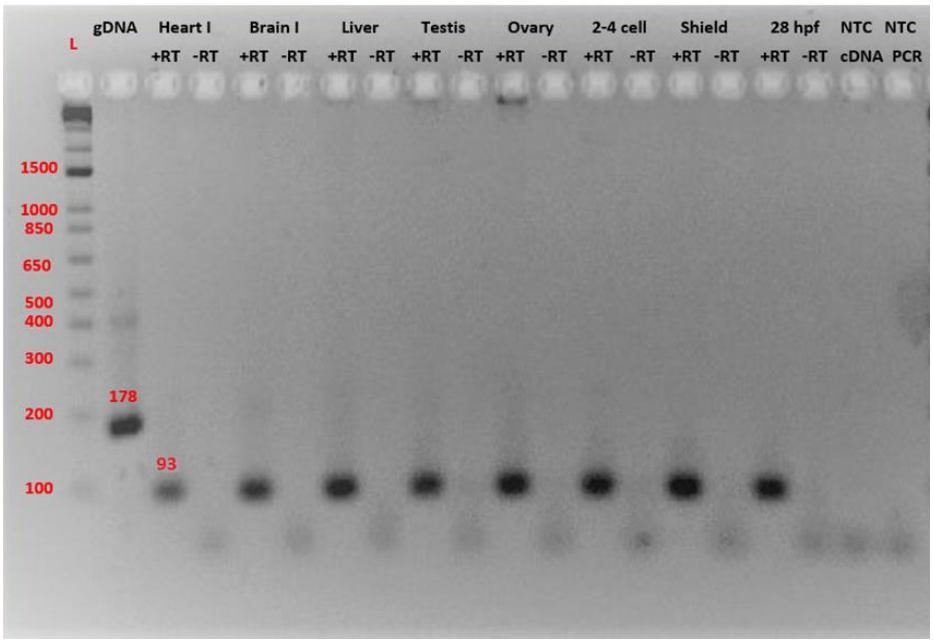


Figure S4. Assessment of genomic DNA contamination in zebrafish RNA samples. *L denotes the ladder; gDNA indicates the sample where genomic DNA was used as a template for the PCR reaction; +RT represents the sample where Reverse Transcriptase was included; -RT indicates the sample where Reverse Transcriptase was not included; NTC cDNA refers to the negative control in which water was added instead of RNA during the reverse transcription reaction; and NTC PCR denotes the negative control for PCR, where water was added instead of cDNA.*

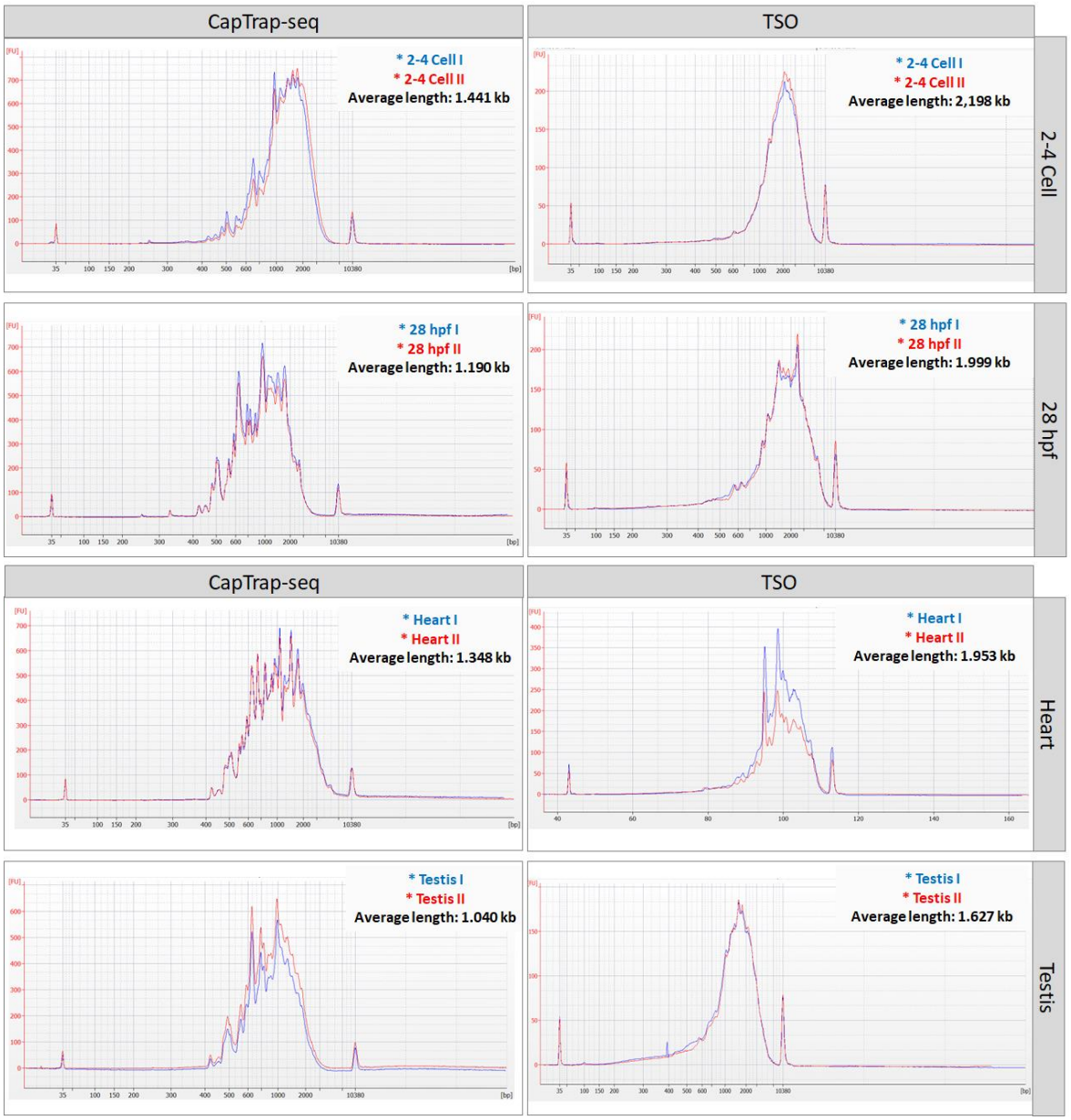


Figure S5. Quality and size distribution of cDNA libraries generated using CapTrap-seq and TSO protocols across various zebrafish samples. *Blue and red profiles indicate two cDNA library replicates, with the average cDNA length shown in the upper right corner.*

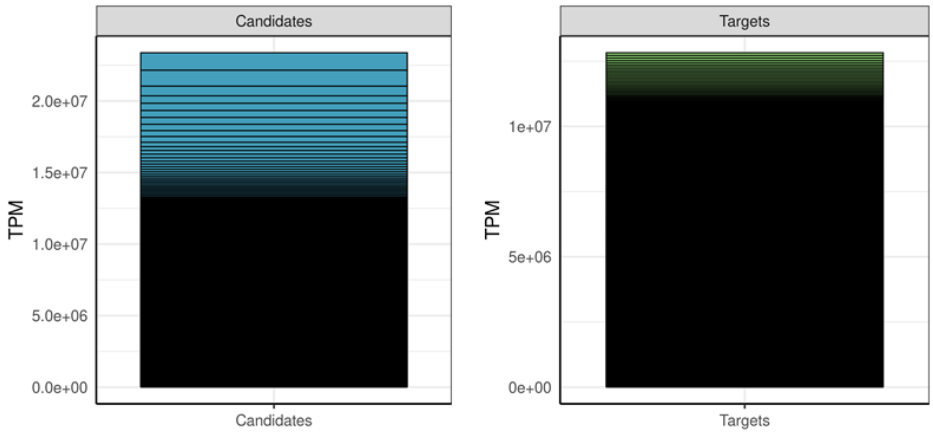


Figure S6. Expression levels for candidate regions (blue) **and targeted regions** (green) **for CapTrap-CLS capture probes design.**

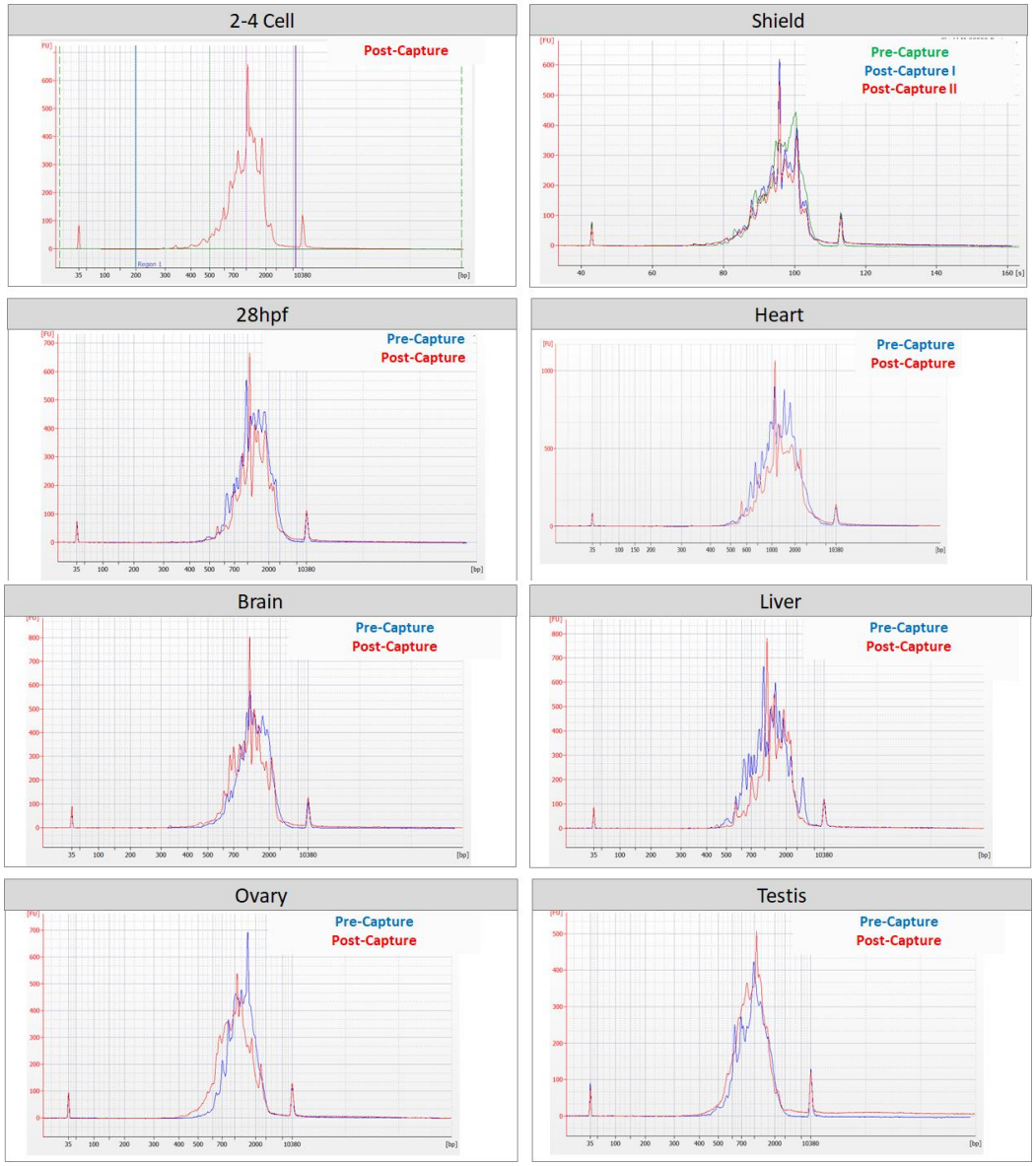
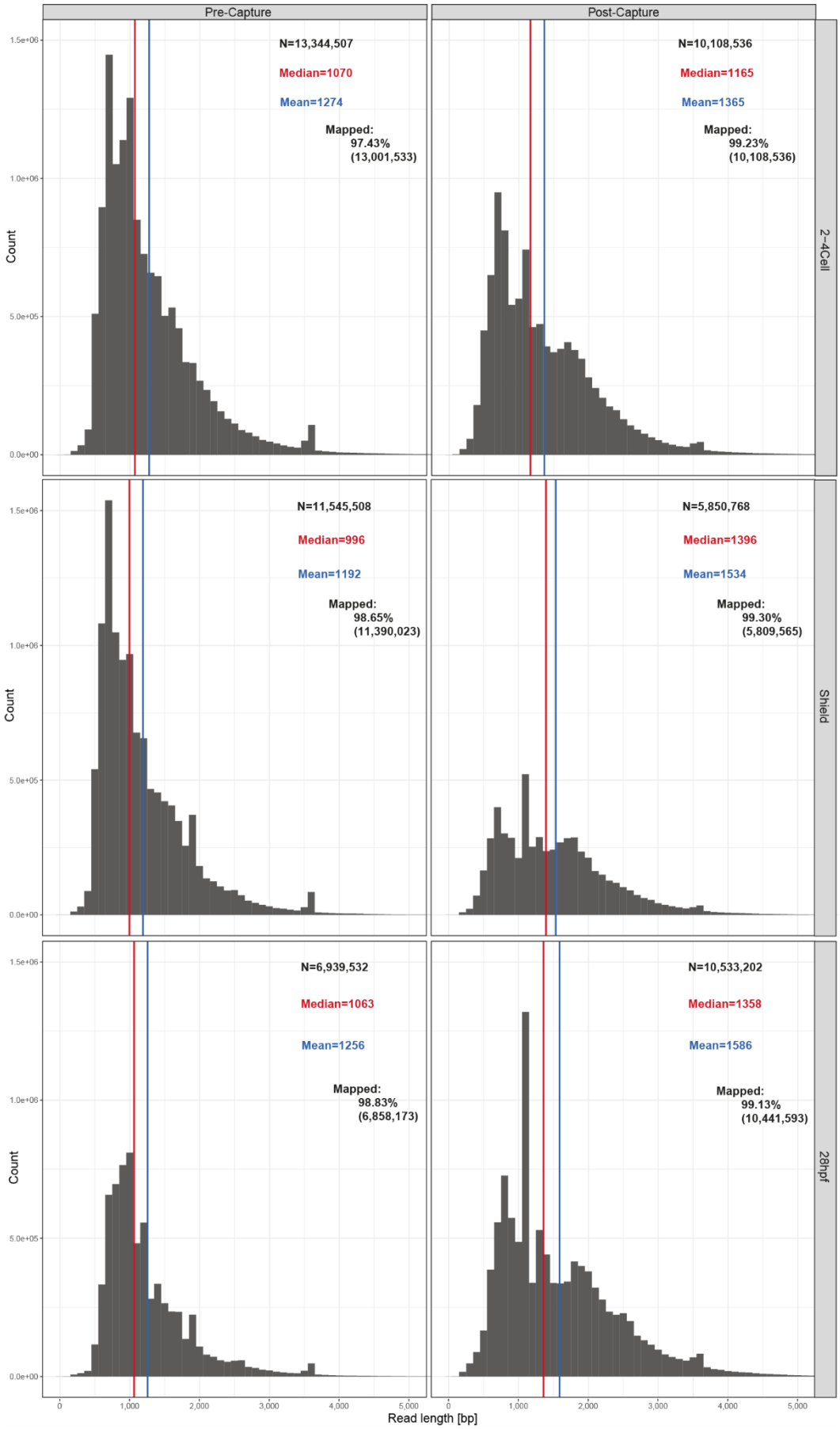
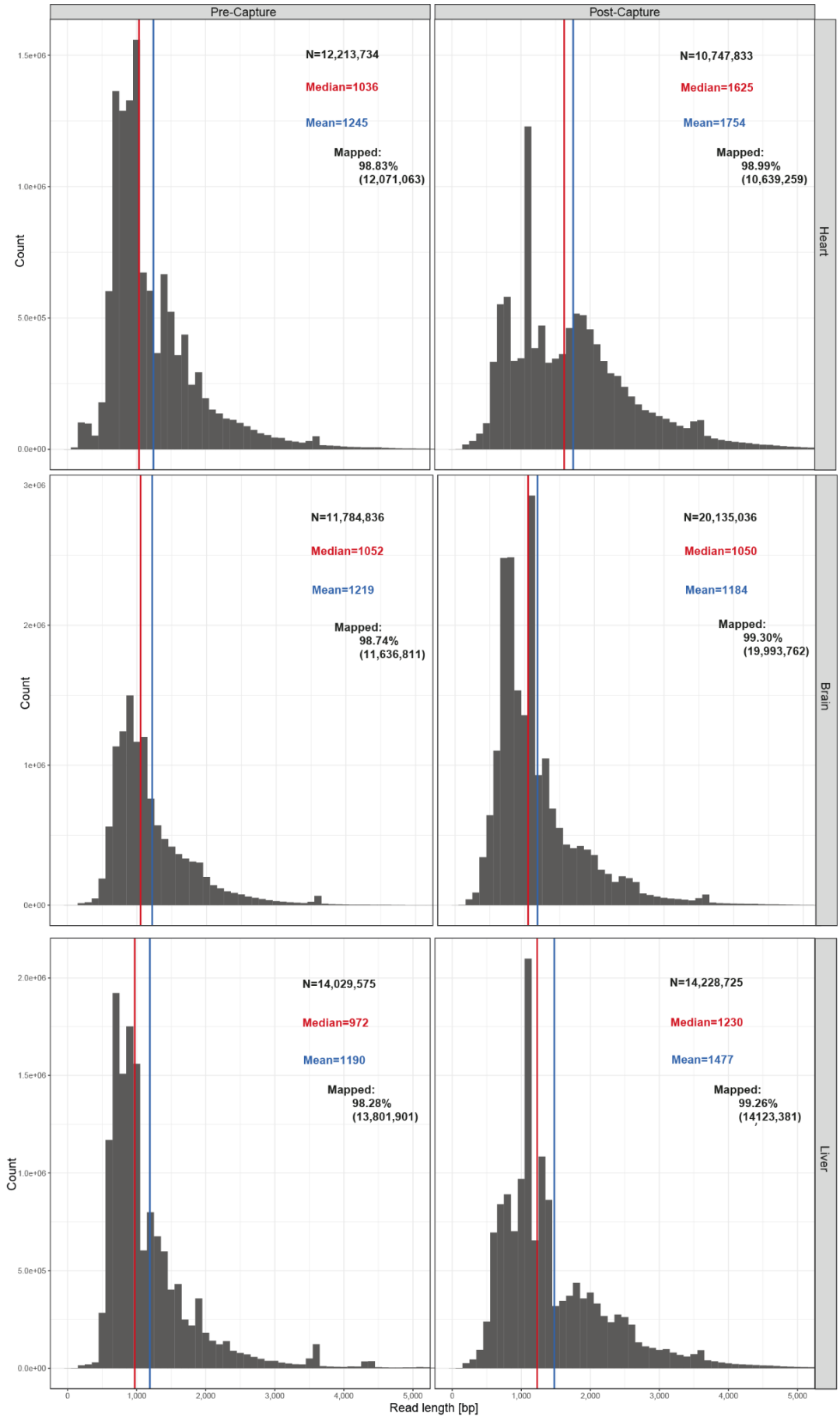


Figure S7. Quality of pre- and post-capture libraries. *Representative electropherogram (Bioanalyzer) illustrating the quality of pre- and post-capture cDNA libraries across various zebrafish samples.*


Pre-Capture
Post-Capture
2-4Cell
Shield
28hpf
Count
Read length [bp]
N=13,344,507
Median=1070
Mean=1274
Mapped:
97.43%
(13,001,533)
N=10,108,536
Median=1165
Mean=1365
Mapped:
99.23%
(10,108,536)
N=11,545,508
Median=996
Mean=1192
Mapped:
98.65%
(11,390,023)
N=5,850,768
Median=1396
Mean=1534
Mapped:
99.30%
(5,809,565)
N=6,939,532
Median=1063
Mean=1256
Mapped:
98.83%
(6,858,173)
N=10,533,202
Median=1358
Mean=1586
Mapped:
99.13%
(10,441,593)
0.0e+00
5.0e+05
1.0e+06
1.5e+06
0
1,000
2,000
3,000
4,000
5,000


Pre-Capture
Post-Capture
Heart
Brain
Liver
Count
Read length [bp]
N=12,213,734
Median=1036
Mean=1245
Mapped:
98.83%
(12,071,063)
N=10,747,833
Median=1625
Mean=1754
Mapped:
98.99%
(10,639,259)
N=11,784,836
Median=1052
Mean=1219
Mapped:
98.74%
(11,636,811)
N=20,135,036
Median=1050
Mean=1184
Mapped:
99.30%
(19,993,762)
N=14,029,575
Median=972
Mean=1190
Mapped:
98.28%
(13,801,901)
N=14,228,725
Median=1230
Mean=1477
Mapped:
99.26%
(14123,381)

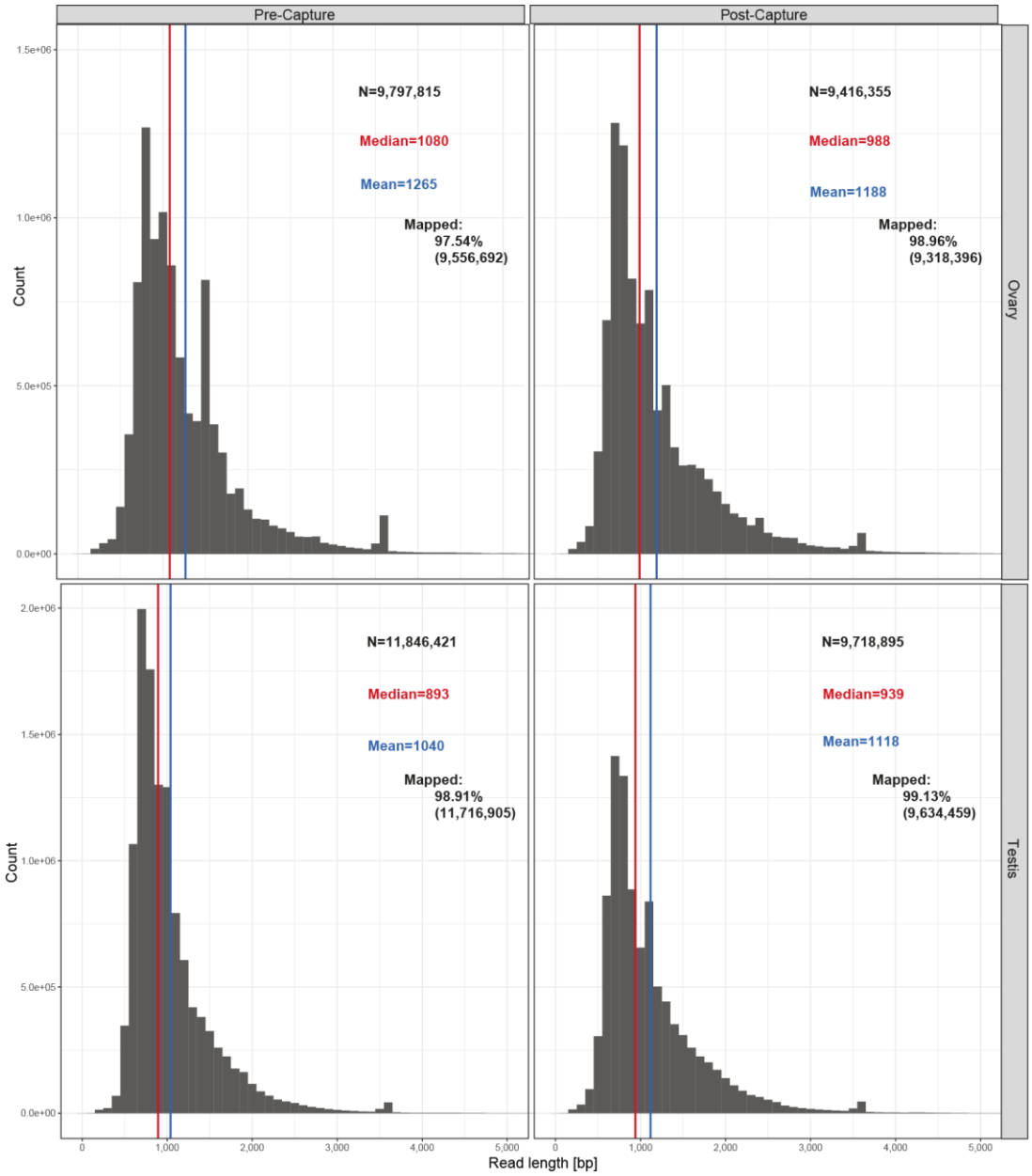


Figure S8. The length distribution of raw long-read ONT reads for pre- and post-capture samples across biological samples. *The total number of reads (N), along with the median and mean read lengths (represented by red and blue vertical lines, respectively), and mapping rate are displayed in the top right corner.*

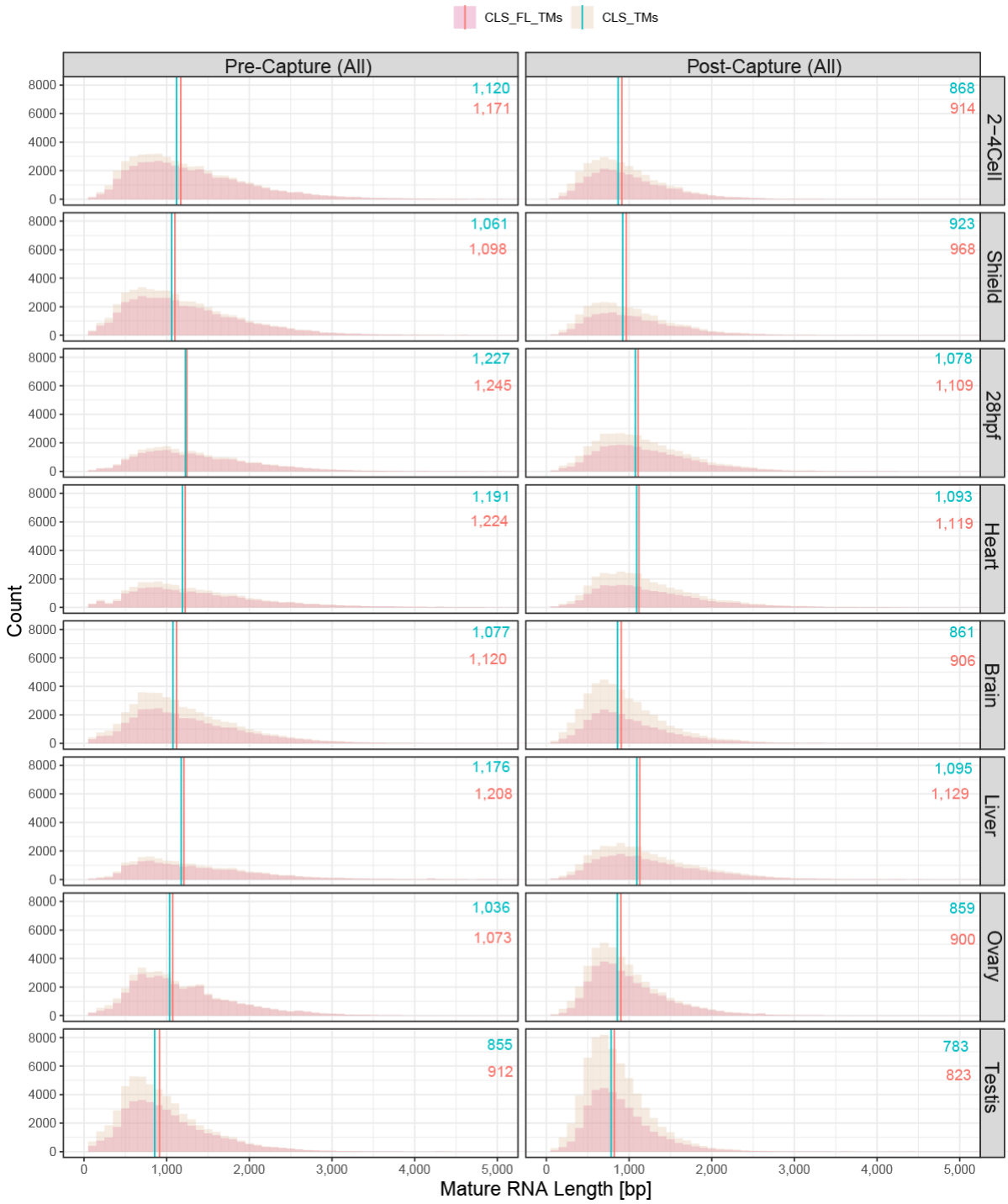


Figure S9. Transcript model length distribution. *Length distribution of complete (peach) and All (beige) transcript models, encompassing all (spliced and unspliced) transcript models. The median read length for all (turquoise) and FL (red) TMs is indicated in the top right corner.*

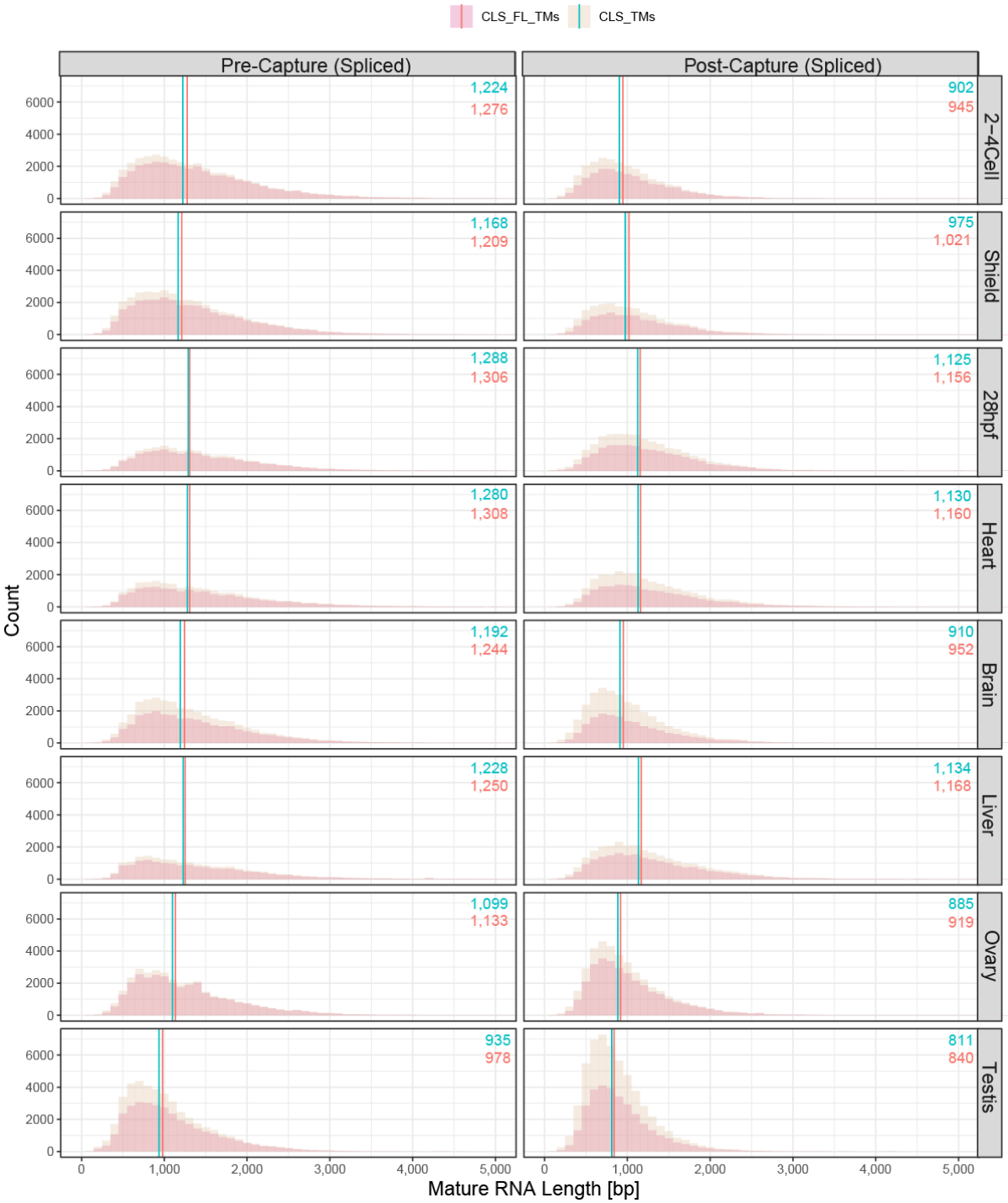


Figure S10. Transcript model length distribution. *Length distribution of complete (peach) and All (beige) transcript models, encompassing spliced transcript models. The median read length for all (turquoise) and FL (red) TMs is indicated in the top right corner.*

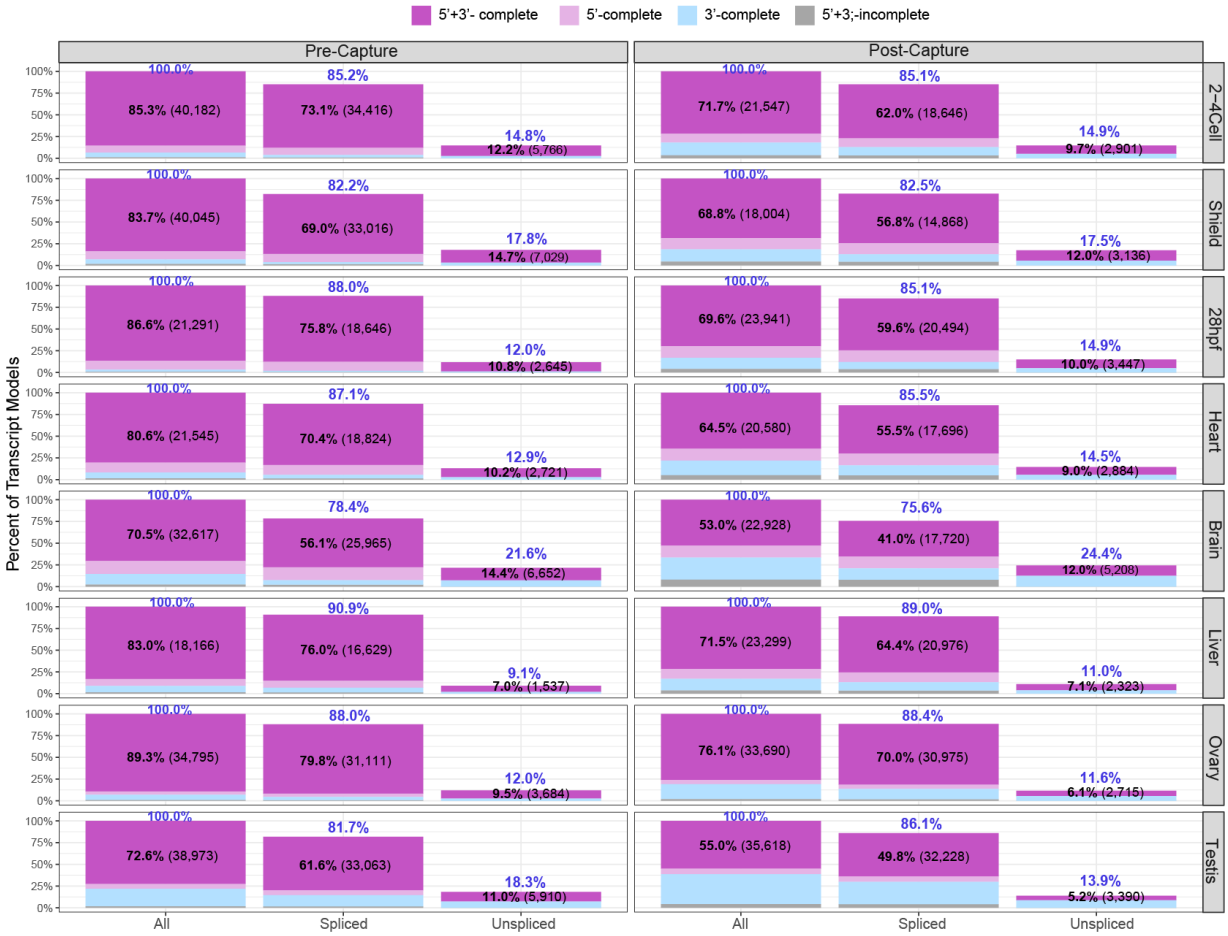


Figure S11. Detection of full-length transcript models (FL-TMs) for pre- and post-capture samples among all, spliced and unspliced TMs. *Four categories of transcript model (TM) completeness are represented: Grey for incomplete TMs, Sky blue for 3'-complete TMs, Light pink for 5'-complete TMs, and Purple for full-length TMs.*

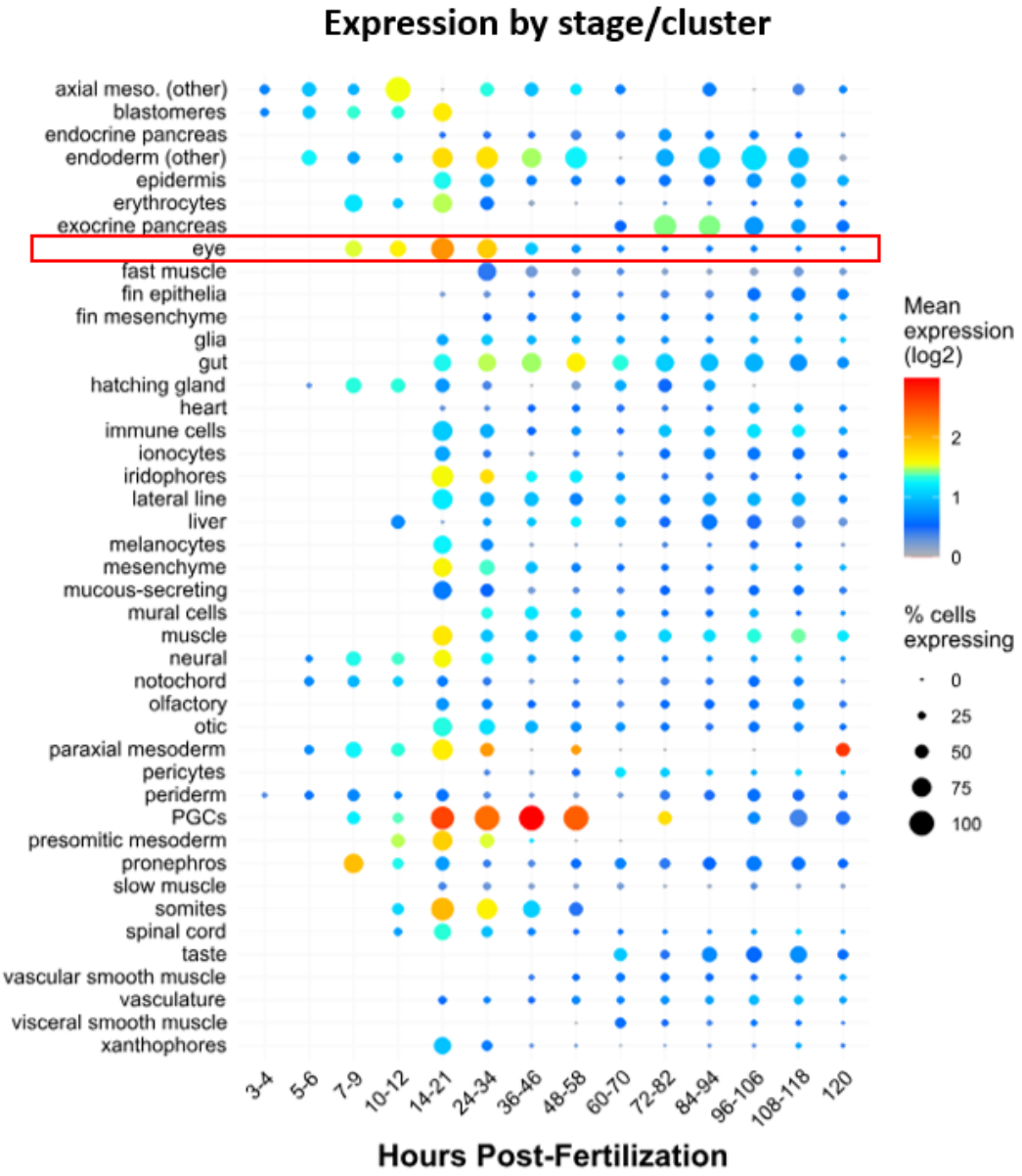


Figure S12. Single-cell expression data for *si:dkey-23i12.5* gene in zebrafish. *Data courtesy Daniocell (Farrell et al., 2018; Sur et al., 2023). The size of the circle corresponds to the percentage of cells expressing the tested gene.*