MASTERARBEIT | MASTER'S THESIS

Titel | Title

Assessing Genome Quality of Molluscan Cephalopod Genomes

verfasst von | submitted by

David Stohlmann BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 831
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Zoologie
Betreut von | Supervisor:	Assoz. Prof. Dr. Oleg Simakov

Table of contents

Acknowledgements

Throughout the work on my master's thesis, I received great support and assistance, which greatly contributed to completing this project and writing my thesis.

First and foremost, I would like to express my sincere gratitude to Oleg Simakov for his invaluable guidance, support and feedback on my work. Thank you for introducing me to the fascinating field of cephalopod genomics and sharing your knowledge. I am also deeply grateful for the opportunity to have been part of your amazing group.

I extend my sincere thanks to Gustavo Sanchez, my co-supervisor, for his patience, expertise, and consistent guidance throughout this project. Despite the fact that we had to work together remotely, you still managed to consistently support me by giving valuable feedback, ideas, and guidance.

I would also like to thank the members of Oleg Simakov's group for their invaluable feedback and encouragement during my project. Your readiness to offer support whenever I needed it was deeply appreciated.

Additionally, I am thankful to my colleagues at the Vienna Zoo for their patience and encouragement. Special thanks go to Roland Halbauer for arousing my interest in cephalopods and aquatic animals in the first place.

Finally, I want to thank my family and friends for their support. To my partner, Denise Piringer, a special thank you for your patience, understanding, and for always being there when I needed someone to talk during challenging times.

This project would not have been possible without all of you, thank you!

Abstract

Over 800 cephalopods, including nautiluses, squids, cuttlefishes, and octopuses, have colonized all marine oceans worldwide. Despite this extensive diversity, just a few cephalopod genomes have been sequenced, the first being the Californian octopus, *O. bimaculoides*, in 2015. The recent increase in the number of cephalopod genomes highlights the need for a standardized assessment approach. Benchmarking Universal Single-Copy Orthologs (BUSCO) is a widely used tool for genome quality assessment, but its current datasets for broader taxa like the eukaryotic, metazoan or molluscan underperformed in cephalopod genomes, yielding high duplications, high fragmentations, high missing genes, and especially low completeness, even in high-quality chromosome-level genomes. We found that the poor performance is due to the complex character of these genomes. To improve this, we have identified a set of 512 cephalopod-specific single-copy orthologous genes from high-quality genomes of *Euprymna scolopes*, *Doryteuthis pealeii*, *Octopus bimaculoides* and *Nautilus pompilius*. This set of genes contains orthologs that are localized on all chromosomes in these species. When our new dataset is used in BUSCO, it yields a better value of completeness that better represents the quality of the available genomes for cephalopods in public databases, making it a valuable addition to the scientific community.

Zusammenfassung

Über 800 Kopffüßer-Arten (Cephalopoda), welche Perlboote, Tintenfische, Sepien und Kraken beinhalten, haben weltweit alle marinen Lebensräume kolonisiert. Trotz dieser großen Vielfalt sind bisher nur wenige Genome von Cephalopoden sequenziert worden, wobei das erste Genom vom Kalifornischen Zweipunktkraken, *Octopus bimaculoides*, aus dem Jahr 2015 stammt. Die in letzter Zeit ansteigenden Zahlen an sequenzierten Genomen, geht mit der Notwendigkeit einer standardisierten Vollständigkeitsbeurteilung einher. „Benchmarking Universal Single-Copy Orthologs“ (BUSCO) ist ein weit verbreitetes bioinformatisches Werkzeug für diese Vollständigkeitsbestimmung. Verfügbare Datensets für eine größere Taxabreite, wie für Eukaryota, Metazoa oder Mollusca, jedoch schneiden unzureichend mit Cephalopodengenomen ab. Das Ergebnis sind hohe Prozentsätze an Duplikationen, Fragmentierungen und fehlenden Genen, trotz der Verwendung von Genomen auf Chromosomlevel. Wir haben herausgefunden, dass diese schlechten Ergebnisse auf die Komplexität der Cephalopodengenome zurückzuführen sind. Um dies zu verbessen, haben wir ein Datenset aus 512 „single-copy“ orthologen Genen erstellt, welche mit Hilfe der hochqualitativen Genome der Arten *Euprymna scolopes*, *Doryteuthis pealeii*, *Octopus bimaculoides* und *Nautilus pompilius*, welcher als Außengruppenvergleich diente, identifiziert wurden. Diese Gene sind über alle Chromosomen dieser Arten gleichmäßig verteilt. Mit der Verwendung unseres Gendatensets in BUSCO konnten wir eine deutliche Verbesserung der Vollständigkeit feststellen, welche die Qualität der veröffentlichten Cephalopodengenome besser repräsentiert. Dies macht unser Gendatenset zu einer wertvollen und unterstützenden Ergänzung für zukünftige wissenschaftliche Forschungen.

1. Introduction

The phylum of Mollusca consists of eight different classes: Caudofoveata, Solenogastres, Polyplacophora, Monoplacophora, Scaphopoda, Bivalvia, Gastropoda and Cephalopoda (Haszprunar & Wanninger 2012). This phylum has the second-highest count of species across all metazoans (Ponder & Lindberg 2008). All mollusks have a shared body plan which represents the phylum's most significant features (Haszprunar & Wanninger 2012).

Mollusks are a highly diverse phylum found across a wide variety of habitats, ranging from oceans to freshwater and terrestrial ecosystems (Ponder & Lindberg 2008). They have even adapted to extreme environments, such as mountaintops, tropical rainforests, deserts, and the high-pressure depths of the deep sea (reviewed in Gomes-dos-Santos *et al*. 2020). Despite sharing a common body plan, mollusks exhibit an immense diversity in body sizes, from microscopic worm-like species to the giant squids (Sigwart *et al*. 2021). Some mollusks, like *Arctica islandica*, are among the longest-living animals in the animal kingdom, with lifespans exceeding 500 years (reviewed in (Gomes-dos-Santos *et al*. 2020).

Mollusks are not only biologically fascinating but also hold economic and medical significance (Sigwart *et al*. 2021). Cephalopods, bivalves, and gastropods, for instance, serve as an essential protein source (Haszprunar & Wanninger 2012), and molluscan shells were historically used as currency as early as the eight century (Colgan 2020). Additionally, mollusks play a vital role in environmental monitoring, aiding the assessment of pollution, ocean acidification, and the impacts of climate change on coastal ecosystems (Farrington *et al*. 2016; Figueras *et al*. 2019; Gazeau *et al*. 2013; Talmage & Gobler 2010; Zhang *et al*. 2012). The vast diversity of molluscan species opens the door to a wide array of research opportunities across different fields.

The largest groups by numbers of species within mollusks are gastropods and bivalves. They represent 96% of the total species, followed by cephalopods as the third most commonly known mollusk class (Brusca 2003 cited in Gomes-dos-Santos *et al*. 2020). There are about 800 extant cephalopod species which colonize all marine waters from pelagic to benthic zones, from shallow waters to the deep sea, as well as from polar regions to the tropics (Jereb & Roper 2005; Packard 1972). Cephalopods are classified into two different groups: the nautiloids, which have retained an external shell reminiscent of ancestral cephalopods, and the coleoids (Albertin & Simakov 2020; Sanchez *et al*. 2018; Young *et al*. 1998). Coleoids comprise two main suborders, the Octopodiformes (octopuses, vampire squids) and Decapodiformes (squids, cuttlefish). They are most easily distinguishable from each other by

their eight or ten arms, respectively, besides other morphological differences (Berthold & Engeser 1987; Boletzky 2003). The divergence between nautiloids and coleoids occurred in the Cambrian (541-485.4 mya), with the latter splitting into Octopodiformes and Decapodiformes about 270 million years ago in the Permian era (Kröger *et al*. 2011; Sanchez *et al*. 2018; Tanner *et al*. 2017). Deeper phylogenetic relationships among coleoid cephalopods are still an ongoing topic and are so far unresolved (Allcock *et al*. 2015; Lindgren & Anderson 2018). Nevertheless, since the division between nautiloids and coleoids, subsequent evolutionary adaptations of coleoid cephalopods have been associated with the adaptation of bony fish. This includes instances of convergent evolution, such as the development of camera-type eyes in both vertebrates and cephalopods, which are considered a morphological novelty among the latter group (Tanner *et al*. 2017). Also, descriptions of new cephalopod species, like three newly described bobtail squids, (Sanchez *et al*. 2019) and the presence of many cryptic speciation, provide more data for future analyses and more insight into the evolution of cephalopods.

Cephalopods have high cognitive capabilities which can be led back to their complex central nervous system. For example, roughly 50 million neurons exist in an octopus's brain (Zullo & Hochner 2011). They have a diverse range of body sizes (3 cm to 7 m, 18 m with arms) (Haszprunar & Wanninger 2012), morphological novelties, such as the light organ of bobtail squids, e.g. *Euprymna scolopes* (see **Fig. 1A**) (Belcaid *et al*. 2019; Ritschard *et al*. 2019), besides the aforementioned camera-type eyes, as well as different lifestyles and behaviors, used for interacting with their biotic and abiotic environment in a myriad of different ways (Hanlon & McManus 2020; Hanlon & Messenger 2018). A fascinating example of communication is the color patterning of the flamboyant cuttlefish *Metasepia pfefferi* (see **Fig. 1B**), which is used for courtship but can also be used as a defense mechanism (Hanlon & McManus 2020; Hanlon & Messenger 2018). When we think of defense mechanisms and cephalopods, the first thing that comes to mind is camouflage, which is a primary defense mechanism where the animal matches perfectly with the surrounding area to avoid detection by a predator or other potential threats. A different color patterning is used as a secondary defense mechanism when camouflage is unsuccessful, using rapid changes to flamboyant colors to scare away predators, as is done by *M. pfefferi* (Hanlon & McManus 2020; Hanlon & Messenger 2018). These remarkable abilities of cephalopods captivate not only morphologists, behavioral/cognitive, developmental, or neurobiologists but recently received increased attention in genomics (Shigeno *et al*. 2015). Researchers in cephalopod genomics are curious what mechanisms shape the cephalopod abilities to gain a deeper insight into their evolution (Ritschard *et al*. 2019).

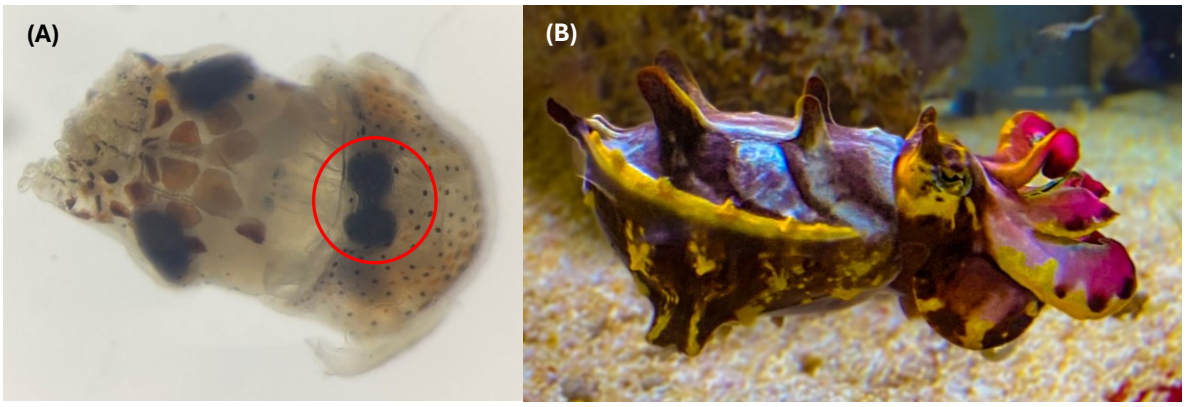

Fig. 1: Picture **(A)** shows the ventral view of *E. scolopes*, with light organ circled in red (Image: David Stohlmann). **(B)** shows the communicative color pattern of *M. pfefferi* (Image: David Stohlmann).

In a 2020 review, which established the state of the situation, the most common mollusk genome assemblies were of gastropods and bivalves and only a small number of cephalopods (Gomes-dos-Santos *et al.* 2020). Four years after, numbers of published cephalopod genome assemblies are increasing. The publication of the first cephalopod genome of *Octopus bimaculoides* by Albertin and colleagues was done in 2015. Thenceforth, 44 genomes of different quality have been assembled. The genome assembly of *Architeuthis dux*, for example, is a scaffold level assembly (da Fonseca *et al.* 2020) and the genome assembly of *Octopus vulgaris* was improved by Destanović *et al.* to chromosome level in October 2023 (Destanović *et al.* 2023).

Generally, such assemblies are on contig-, scaffold- or chromosome-level. A genome assembly starts with extracting the DNA or RNA from tissues, followed by different sequencing methods. Briefly, short or long reads can be generated by various sequencing methods. For example, Illumina sequencing can be used to get short sequencing reads (50 – 300 base pairs) or Pacific Biosciences (PacBio) technologies can produce long sequencing reads (10,000 – 15,000 bp). These sequencing reads are further used to annotate the genome and are assembled to contigs in further steps. During scaffolding, those contigs are put together to even longer fragments, the so-called scaffolds. In the last step, those scaffolds can be assembled to chromosome level. These different assembly and annotation steps can be done using various programs choosing whichever suits the specific analyses' purpose best. This is of essential importance as misassembles of repetitive elements due to unsuitable software can impact the final quality of genome assemblies (Dominguez Del Angel *et al.* 2018), underlining the importance of having suitable programs available.

Cephalopod genomes are generally large and highly repetitive, ranging from 2.5 – 5 giga base pairs (Gb). For example, the genome size of the bivalve *Crassostrea gigas* and gastropod *Pomacea maculata* is 545 mega base pairs (Mb) and 414.8 Mb, respectively (reviewed in

Gomes-dos-Santos *et al*. 2020). In comparison, cephalopods can even have a larger genome size than humans with a size of ~3 Gb (Sohn & Nam 2016). This high number of base pairs makes it difficult to sequence and assemble these genomes, which is one of the reasons for the late publication of the first cephalopod genome compared to other metazoan or even molluscan genomes (Albertin *et al*. 2015; Kim *et al*. 2018; Schmidbaur *et al*. 2022). All available cephalopod genomes enable a deeper insight into the genomic background of cephalopod evolution and the emergence of novelties (Albertin *et al*. 2015; Kim *et al*. 2018; Schmidbaur *et al*. 2022). However, while the majority of genome sizes are within the range above of 2.5 to more than 5 Gb, there are exceptions. The genome size of *Idiosepius* is smaller at 2.1 Gb (Albertin *et al*. 2012). A possible explanation for the enormous sizes of these genomes has been proposed to be whole-genome duplication within cephalopods (Bonnaud *et al*. 2004; Hallinan & Lindberg 2011). Still, other studies do not support this idea and instead assume whole-genome rearrangements are responsible (Albertin *et al*. 2015; Belcaid *et al*. 2019). Rearrangements can happen due to whole chromosomal reorganizations but can also occur locally in a way that changes the order of genes. Cephalopod genomes also show noteworthy gene family expansions, and together with rearrangements, represent significant characteristics to their genomes (reviewed in Albertin & Simakov 2020; Ritschard *et al*. 2019).

A high proportion of cephalopod genomes consist of repetitive elements (Albertin *et al*. 2012, 2015; Kim *et al*. 2018; Schmidbaur *et al*. 2022), for example, almost 50% of repetitive regions cover the draft genome of *A. dux* (da Fonseca *et al*. 2020), which is a reason for the large cephalopod genome size differences (Albertin & Simakov 2020). Ultimately, all these genome characteristics of cephalopods likely contribute to their high morphological diversity as well as novelties (Ritschard *et al*. 2019).

As previously addressed, genome sizes between species, as well as the assembly levels, differ greatly between different publications, which makes it difficult to compare genome quality among existing assemblies. Back in 2007, a bioinformatics tool called CEGMA (Core Eukaryotic Genes Mapping Approach) was published by Parra and colleagues, which was the first tool capable of identifying highly conserved genes in a new genome assembly (Parra *et al*. 2007). Due to resource problems and an upcoming new tool, CEGMA had to be shut down in 2015 (Bradnam 2015), and shortly after, a new tool was published by Simão and colleagues called Benchmarking Universal Single-Copy Orthologs (BUSCO, Simão *et al*. 2015) which is available online to this day. As the name already suggests, it uses single-copy orthologous genes to measure genomic data completeness.

Orthologous genes are derived by speciation events, in contrast to paralogous genes which can be traced back to duplications (Gabaldón & Koonin 2013) and used for identifying ancestral genes that were present in the last common ancestor of compared species (Dolinski & Botstein 2007; Koonin 2005; Sonnhammer & Koonin 2002). Orthologous genes have biologically identical functions in different species, whereas paralogous genes' functions normally diverge after duplication events (Koonin 2005; Studer & Robinson-Rechavi 2009). Orthologs are the most similar ones across different species, regarding sequences, structures and functions (Gabaldón & Koonin 2013) and can be used to describe differences as well as similarities of genomes of various species (Dolinski & Botstein 2007; Koonin 2005; Sonnhammer & Koonin 2002).

BUSCO (Simão *et al*. 2015) provides insight into how many orthologs are present in the genome of interest and how many are single-copy, duplicated, or fragmented orthologs and which are missing (Simão *et al*. 2015). Currently, there are many different lineages available, such as metazoa, eukaryota, fungi, bacteria, and many more. A complete list of all lineages can be viewed online at List of BUSCO.v4 lineages (ezlab.org). Unfortunately, none of them are suitable for cephalopod genomes, especially the metazoan, eukaryotic or even the molluscan lineage. The current problems with the usage of assessing cephalopod genome completeness are low BUSCO scores. To determine if an orthologous gene is complete, duplicated, fragmented, or missing, two parameters are significant: the score and the length cutoff parameter. The former is defined as the 90% minimum bit score value of an orthologous group (orthogroup), whereas the length cutoff parameter is fulfilled when the length of an ortholog is within two times the standard deviation of the orthogroup's mean length. A complete orthologous gene is classified as a single-copy ortholog when fulfilling the score cutoff parameter whereas if multiple matching genes fulfill both parameters it is a duplicated gene. If the ortholog length does not achieve the length cutoff it is classified as fragmented and if the score nor the length cutoff is fulfilled it is a missing gene (Simão *et al*. 2015). Generally, the different BUSCO categories provide insight into how complete a genome of interest is regarding a core set of orthologous genes. In a 2022 study of *Euprymna berryi* genome (5.6 Gb, N50 value of 113.96 Mb) the authors received a completeness of only 85%, and almost 10% were missing genes with the metazoan dataset (Gavriouchkina *et al*. 2022). The newly published genome assembly of *O. vulgaris* (2.8 Gb, N50 value of 118.9 Mb) revealed a completeness of 92% and 5% missing genes with the metazoan dataset. In this study the authors used the molluscan dataset as well, but the results were even worse with only 86% completeness and 10% missing genes (Destanović *et al*. 2023).

Given the complexity of assembling cephalopod genomes and the limited availability thereof, providing a core set of single-copy cephalopod-specific orthologous genes for the bioinformatics tool BUSCO (Simão *et al*. 2015) is long overdue. The goal is to establish this set of orthologs, because currently used BUSCO datasets are not suitable for cephalopods. For this purpose, to generate a cephalopod-specific BUSCO lineage, already published high-quality (chromosome-scale) genome assemblies of four different species were used. The aim of the present study is to develop two distinct databases, each created in different ways, ultimately selecting the one that fits the most. This includes having long orthologous genes with complete open reading frames (ORF) and a gene distribution across all chromosomes. The first database underlies gene model-based data, the second database will incorporate RNA-sequence data. In the end, a cephalopod-specific BUSCO dataset is necessary for accurately assessing genome completeness in cephalopods, as it is adjusted to the unique characteristics of cephalopod mollusks. This specialized dataset will also improve future genome assembly and annotation approaches by providing a more precise benchmark for these processes. Additionally, the final dataset will be used to show how the core set of cephalopod-specific orthologs can be used in other analyses, like chromosomal evolution analyses. The final cephalopod-specific database holds the promise to be a valuable benefit, offering a stronger foundation for future research in cephalopod evolution.

2. Material and methods

2.1. Species selection

For this study, genome assemblies and RNA-sequence (RNA-seq) data of three coleoid cephalopod species and one nautiloid species were employed. They were selected based on their quality (chromosome-scale) genome assemblies and their diverse distribution across the cephalopod evolutionary tree. These species included one member of the Sepiolida lineage, the Hawaiian bobtail squid *Euprymna scolopes*, one member of the Myopsida lineage, the Boston market squid *Doryteuthis pealeii*, one octopod species, the California two-spot octopus *Octopus bimaculoides* and one nautiloid species, the chambered nautilus *Nautilus pompilius*.

E. scolopes (see **Fig. 2A**) is a decapodiform, sepiolid bobtail squid native to the coastal regions of the Hawaiian archipelago. This nocturnal species forms a symbiotic relationship with the bacterium *Vibrio fischeri* (McFall-Ngai 2008). The first genome assembly of *E. scolopes* was published in Belcaid and colleagues in 2019 and subsequently refined in Schmidbaur *et al.* in 2022. The genome size is approximately 5.1 Gb, with an N50 value of 120.3 Mb, and is organized into 46 chromosome-scale scaffolds (2N = 92 karyotype), with 94.23% of the genome contained within these scaffolds.

D. pealeii (see **Fig. 2B**) is a decapodiform, loliginid squid found on the continental shelf and in upper slope waters of the northwestern Atlantic (Jereb & Roper 2005), whose genome was first assembled in Albertin *et al.* in 2022 with an assembly size of approximately 4.6 Gb, with an N50 value of 107.4 Mb. The genome comprises 46 chromosome-scale scaffolds (2N = 92 karyotype), with 98.36% of the genome within these scaffolds.

O. bimaculoides (see **Fig. 2C**) is an octopodiform species found in intertidal to subtidal zones of the northeastern Pacific (Jereb *et al.* 2005). Albertin *et al.* (2015) first published the genome and updated it in 2022. The genome assembly is approximately 2.3 Gb, with an N50 value of 96.9 Mb. *O. bimaculoides* is organized into 30 chromosome-scale scaffolds (2N = 60 karyotype), with 95.46% of the genome within these scaffolds.

N. pompilius (see **Fig. 2D**) is one of the few cephalopod species with an external shell (Kröger *et al.* 2011) and inhabits the Indo-Pacific region at depths between 100 and 600 meters (Vandepas *et al.* 2016). Zhang *et al.* (2021) first published the genome, which is approximately 785.15 Mb, with an N50 value of 32.5 Mb. *N. pompilius* has 26 chromosome-scale scaffolds (2N = 52 karyotype), with 93.82% of the genome contained within these scaffolds.

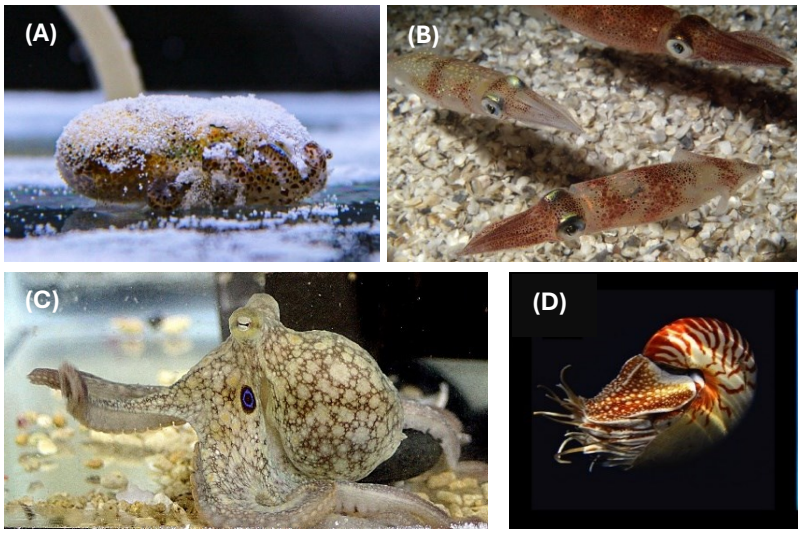

Fig. 2: The four species used to generate the cephalopod-specific core set of orthologous genes. **(A)** *Euprymna scolopes* (Image: David Stohlmann), **(B)** *Doryteuthis pealeii* (Image: Albertin *et al*. 2022), **(C)** *Octopus bimaculoides* (Image: An Octopus Emerges | Marine Biological Laboratory (mbl.edu)) and **(D)** *Nautilus pompilius* (Image: Vandepas *et al*. 2016).

2.2. Customized BUSCO lineage generation

2.2.1. General procedures

For the cephalopod-specific dataset, adjusted according to Benchmarking Universal Single-Copy Orthologs (BUSCO v5.4.6) (Simão *et al*. 2015), the four species previously introduced were utilized.

Initially, OrthoFinder 2.5.5 (Emms & Kelly 2015) was employed to identify conserved single-copy orthologous genes across these species. Subsequently, multiple sequence alignment was performed using MAFFT 7.520 (Katoh & Standley 2013). Hidden Markov Models (HMMs) were then utilized via HMMER 3.4 (Eddy 2011) to search for homologous sequences, align them, and derive orthologous gene profiles.

HMMER 3.4 facilitated the setting of two parameters: score and length cutoff. These parameters played a significant role in identifying true orthologous genes and distinguishing them into single-copy, duplicated, fragmented or missing orthologs (see **Fig. 3B**). The score cutoff parameter is defined as the 90% minimum bit score value of an orthologous group (orthogroup). The length cutoff was determined such that the mean length of an ortholog had to fall within two standard deviations of the mean length of the orthogroup. An ortholog was classified as a single-copy ortholog if it fulfilled both score and length cutoffs. If multiple matching genes met these criteria, they were classified as duplicated. Orthologs failing to meet the length cutoff were classified as fragmented, while those not meeting either cutoff were classified as missing.

BUSCO v5.4.6 typically uses default settings for maximum intron size (130,000 bp) and maximum sequence length (160,000 bp) for other datasets, such as the widely used metazoan or eukaryotic dataset. However, we estimated the intron size distribution within the four cephalopod genomes: *E. scolopes* 93,502 bp, *D. pealeii* 321,820 bp, *O. bimaculoides* 86,956 bp and *N. pompilius* 806,852 bp. Based on these estimates, we increased the maximum intron size in the customized BUSCO lineage to improve the reliability of completeness for the orthologous genes, while retaining the default setting for the maximum sequence length.

Following these adjustments, BUSCO v5.4.6 identified some missing genes. A manual examination revealed that these genes were not true orthologs, as they did not meet the score cutoff parameter, and therefore were subsequently removed from the final list. Finally, Miniprot v0.12 (Li 2023) was applied in conjunction with BUSCO v5.5.0 to obtain the final core set of cephalopod-specific single-copy orthologous genes (see **Fig. 3**).

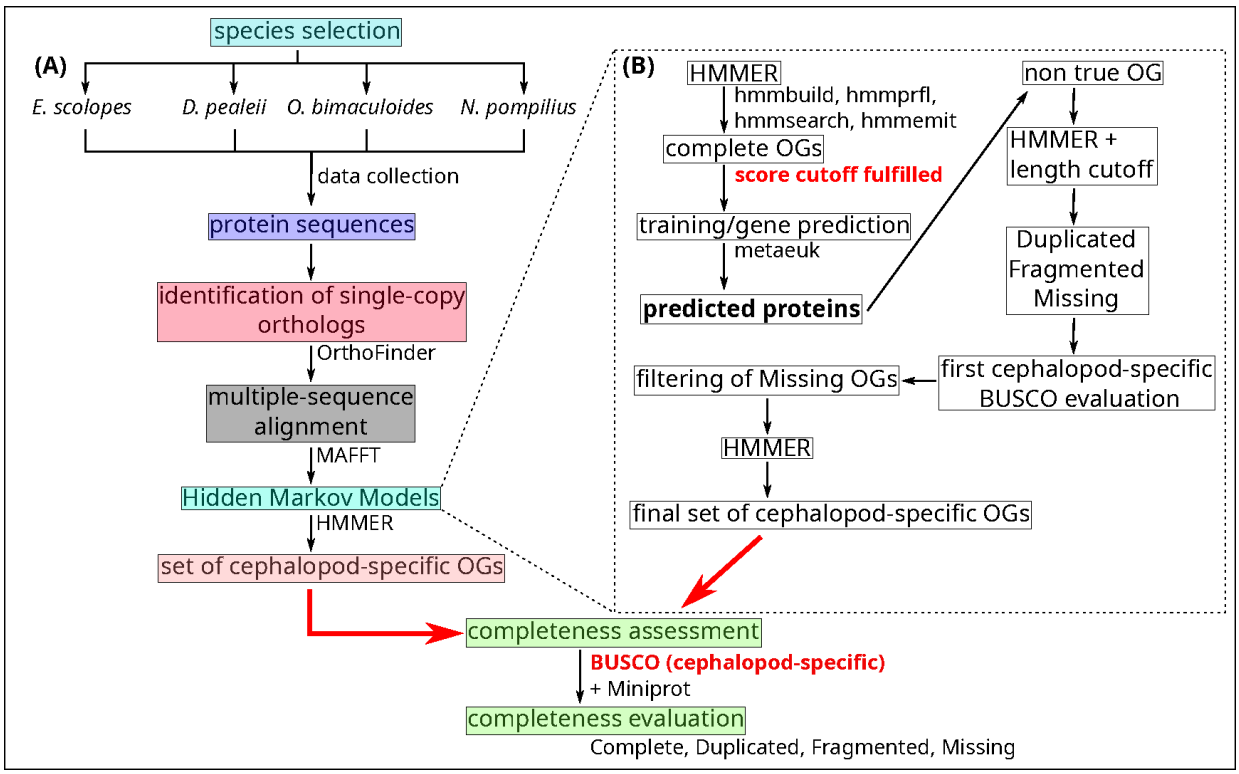

Fig. 3: Pipeline of the generation of the cephalopod-specific BUSCO dataset. **(A)** shows the generally necessary procedures and **(B)** the Hidden Markov Models in detail.

2.2.2. Cephalopod-specific BUSCO lineage version 1

To run OrthoFinder 2.5.5, we utilized the protein sequences previously published for each species. Using the identified single-copy orthologs, we constructed the first version of the cephalopod-specific BUSCO dataset. The methodology for this construction is detailed in

section 2.2.1., General procedures. The gene annotation for this version relied on the published protein data, which is based on gene modeling.

2.2.3. Cephalopod-specific BUSCO lineage version 2

The MIKADO v2.3.4 pipeline (Venturini *et al*. 2018) was used in this analysis to generate the protein sequences necessary for running OrthoFinder 2.5.5. The MIKADO v2.3.4 pipeline was chosen for its ability to integrate multiple evidence to produce transcript annotations from various multiple RNA-seq assemblies. Additionally, it facilitated the selection of genes with open reading frames (ORFs) (Venturini *et al*. 2018), providing a solid reference for evaluating completeness.

Quality control of the RNA-seq data was performed using *fastp* (Chen *et al*. 2018). Subsequently, the cleaned RNA-seq reads were mapped to the reference genome using two different mapping programs: Spliced Transcripts Alignment to a Reference (STAR version 2.7.11a) (Dobin *et al*. 2013) and Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT2 version 2.2.1) (Kim *et al*. 2019). These mapping programs produced compressed binary versions of sequence alignment map (SAM) files, known as BAM files. The BAM files were then used to assemble RNA-seq alignments per organ into transcriptomes with the network flow algorithm StringTie v2.2.1 (Pertea *et al*. 2015), which generated files in gene transfer format (GTF).

Multi-sample transcriptomes for each mapping program (STAR 2.7.11a and HISAT2 2.2.1) were subsequently generated using the meta-assembly method Transcriptome Assemblies Combined to One (TACO, version 0.7.3) (Niknafs *et al*. 2017). TACO 0.7.3 also provided GTF files, which were then used as input for the MIKADO v2.3.4 pipeline.

In addition to the output of TACO 0.7.3, the genome-independent transcriptome analysis workflow Trinity 2.15.1 (Haas *et al*. 2013) was used to perform *de novo* transcriptome assemblies from RNA-seq data. The resulting transcript assemblies were then aligned using the Genomic Mapping and Alignment Program (GMAP), because of its accuracy in detecting splice sites (Wu & Watanabe 2005). The GMAP alignment produced generic feature format (GFF) files, which were essential for subsequent steps in the MIKADO v2.3.4 pipeline.

For the MIKADO v2.3.4 pipeline, transcriptomes generated by the TACO 0.7.3 method were given priority over those produced by the Trinity 2.15.1 workflow. Before running the MIKADO v2.3.4 pipeline, reliable splicing junctions were identified using Portcullis 1.2.4 (Mapleson *et al*. 2018). BAM files generated by STAR 2.7.11a (Dobin *et al*. 2013) were preferred for this step due to their higher unique alignment rate compared to HISAT2 2.2.1. Portcullis 1.2.4 provided

output in browser extensible data (BED) format, which was necessary for the MIKADO v2.3.4 pipeline.

Following the preparations, the MIKADO v2.3.4 pipeline was run to generate the final high-quality transcriptomes (see **Fig. 4**). A detailed description of the MIKADO pipeline is given in https://doi.org/10.1093/gigascience/giy093.

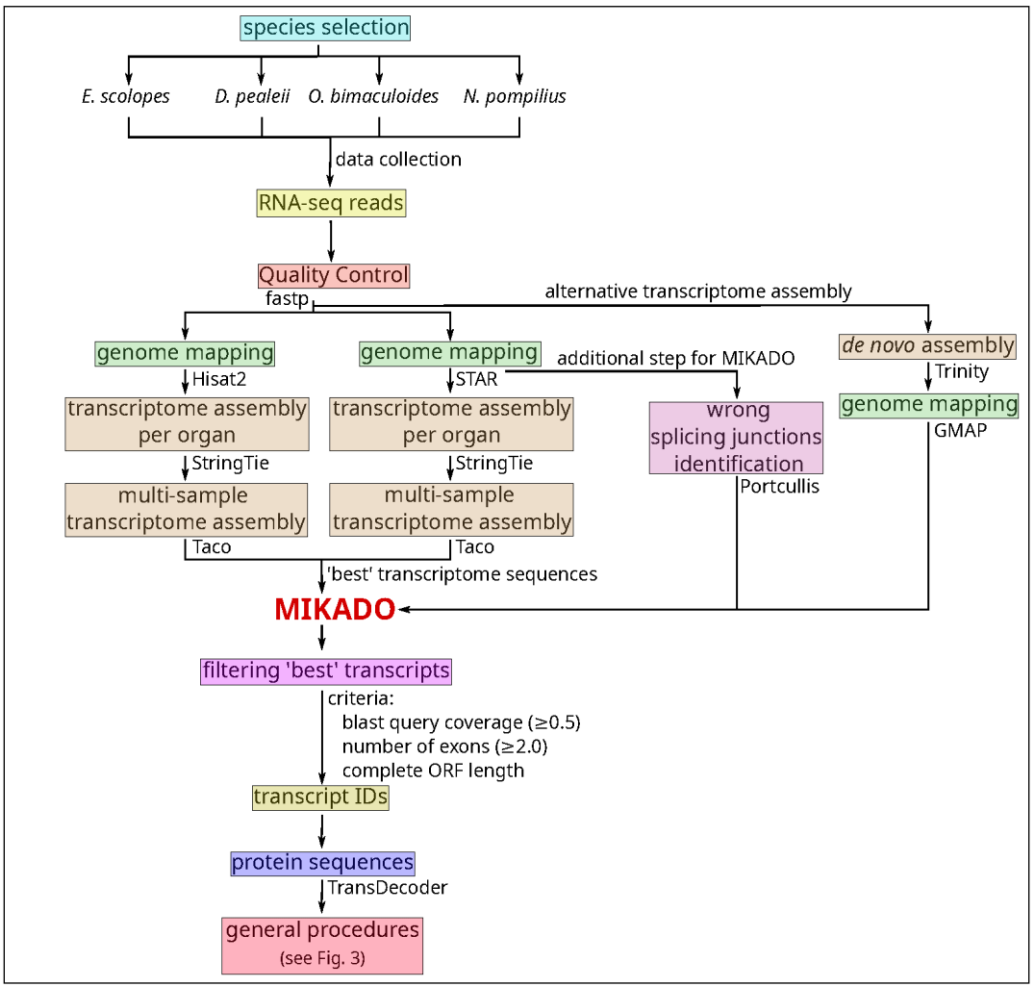

Fig. 4: Preparation pipeline for generating the cephalopod-specific BUSCO lineage version 2 using the MIKDO v2.3.4 pipeline.

After obtaining the final transcriptomes from the MIKADO v2.3.4 pipeline, a filtering process was conducted to select a complete transcriptome reference. The filtering criteria included a Basic Local Alignment Search Tool (BLAST) query coverage of ≥ 0.5, a minimum of two exons, and a complete length of ORFs. Transcriptomes fulfilling these criteria were listed by their ID.

This list was then used to identify the corresponding high-quality protein sequences that were then retrieved using the peptide file generated by TransDecoder v5.7.1, which is able to identify coding regions within transcript sequences (Haas 2023).

The selected protein sequences were then used as input for OrthoFinder 2.5.5 to identify single-copy orthologs among the four species of interest. The remainder of the construction for this cephalopod-specific BUSCO dataset followed the procedures described in section 2.2.1., General procedures.

2.2.4. Coleoid-specific BUSCO lineage

Given the diverse adaptations among various species within the mollusk class of cephalopods, the shelled nautiloid *N. pompilius* was replaced by a second octopod species, the common octopus *Octopus vulgaris* (see **Fig. 5**). This substitution was made to explore potential variations or changes in the set of orthologous genes and to discover any differences. The other cephalopod species – *E. scolopes*, *D. pealeii* and *O. bimaculoides* – were still included.

This customized BUSCO lineage was generated in the same manner as the cephalopod-specific BUSCO lineage version 2 (see section 2.2.3., Cephalopod-specific BUSCO lineage version 2).

Fig. 5: The common octopus *O. vulgaris*, used among others to generate the core set of coleoid-specific orthologous genes (Image: David Stohlmann).

2.3. Length of orthologous genes

To focus on the quality of gene sizes, the analysis aimed to determine which cephalopod-specific BUSCO lineage version produced larger orthologous genes. Therefore, the mean and median sequence lengths of single-copy orthologs were calculated and compared specifically for cephalopod-specific BUSCO lineage version 1 and version 2. No comparison with the coleoid-specific BUSCO lineage was conducted, as the primary focus was on the core set of cephalopod-specific single-copy orthologs.

The FASTA files of single-copy orthologs provided by OrthoFinder 2.5.5 were used to calculate the mean sequence length of each orthogroup for both the cephalopod-specific BUSCO lineage version 1 and 2 gene sets. Additionally, the median length of each orthogroup was calculated.

The plot was created using the Python environment Spyder IDE 5.5.3 (Raybaut & Cordoba 2009). The version of python used during the study was Python 5.15.10 (Python Software Foundation 2001). The libraries pandas (The pandas development team 2024) and matplotlib.pyplot (Hunter 2007) were used for data manipulation and graph generation, respectively.

2.4. BUSCO lineage comparisons with other datasets

To identify overlapping genes among the metazoan (metazoa_odb10), eukaryotic (eukaryota_odb10) and molluscan (mollusca_odb10) BUSCO lineages in relation to cephalopod-specific and coleoid-specific BUSCO datasets, we generated a Venn diagram. Initially, protein sequences of *E. scolopes*, *D. pealeii*, *O. bimacuoloides* and *N. pompilius*, derived from the cephalopod-specific BUSCO lineage version 2, were analyzed using the protein mode of BUSCO v5.5.0 with the metazoa_odb10, eukaryota_odb10 and mollusca_odb10 lineages to identify single-copy orthologous genes.

Subsequently, sequences of each single-copy ortholog per species were concatenated into individual FASTA files, resulting in four FASTA files per species – one per lineage – culminating in a total of 16 FASTA files. These files were processed using OrthoFinder 2.5.5 to identify homologous single-copy orthologs across these four species for each BUSCO lineage.

Post-OrthoFinder analysis, the Orthogroups.GeneCount.tsv was used, which contains information on the occurrence of orthogroups across the input files, including the number of genes in each orthogroup. This data was then utilized to generate a Venn diagram using R version 4.4.1 (R Core Team 2024) and RStudio version 2024.4.2.764 (Posit team 2024), employing the VennDiagram version 1.7.3 (Hanbo Chen 2011) as well as the dplyr version 1.1.4 (Wickham *et al.* 2023) packages.

An analogous procedure was conducted for the coleoid-specific BUSCO dataset, substituting *N. pompilius* with *O. vulgaris*.

Additionally, we generated a bar chart using the Python environment Spyder IDE 5.5.3 to compare the BUSCO results between the cephalopod-specific BUSCO lineage version 2, metazoan (metazoa_odb10), eukaryotic (eukaryote_odb10), and molluscan (mollusca_odb10)

datasets, with data extracted from each BUSCO output. For data manipulation, the libraries pandas (The pandas development team 2024) and matplotlib.pyplot (Hunter 2007) were utilized.

Furthermore, we used four additional cephalopod species to evaluate the performance of our dataset: *Sepioteuthis lessoniana*, *Sepiola atlantica*, *Thysanoteuthis rhombus*, and *Octopus sinenis*. These BUSCO results were then used to create another bar chart, using the Python environment Spyder IDE 5.5.3 and the libraries pandas (The pandas development team 2024) and matplotlib.pyplot (Hunter 2007).

2.5. Cephalopod- and coleoid-specific gene orthology

To determine the localization of cephalopod- as well as coleoid-specific orthologs in homologous chromosomes across different species, we utilized the R package macrosyntR (El Hilali & Copley 2023) to generate an Oxford dot plot. MacrosyntR is designed to identify genome-wide synteny conservations. Prior to running macrosyntR, the input was prepared, requiring three specific files. The first file was a table with two columns listing orthologous gene names of the compared species, which was obtained from OrthoFinder 2.5.5. The second and third file consisted of BED files containing sequence names and genomic coordinates for all orthologs of species one and species two, respectively, ensuring the names matched those in the ortholog table.

Once the input data was prepared, macrosyntR was employed using R version 4.4. and RStudio version 2024.4.2.764. The R packages ggplot2 version 3.5.1 (Wickham 2016) and devtools version 2.4.5 (Wickham *et al.* 2022) were also utilized to generate the Oxford dot plot as well as Inkscape v.1.3.2 (Harrington *et al.* 2003) for final graph manipulations.

2.6. Gene ontology enrichment analysis

For this part of the study, we used the unique cephalopod- and coleoid-specific orthologs identified while comparing the different available BUSCO lineages (see section 2.4., BUSCO lineage comparisons with other datasets). As a first step, these genes were subjected to protein-to-protein BLAST (blastp) using the sensitive mode of DIAMOND (Buchfink *et al.* 2015), which, although slower, offers greater accuracy. The SwissProt database served as the reference database. For each query sequence, the top hit was selected with an e-value threshold of $1x10^{-25}$. The ID of each top blastp hit was then compiled and used for Gene Ontology (GO) enrichment analysis (Thomas *et al.* 2022), which was conducted using the web-based tool available at GO enrichment analysis (geneontology.org). GO is used to identify

protein classes, biological processes, molecular functions, or also cellular components of specific gene sets.

The GO term analysis results were downloaded for both sets of orthologs, and further processed in the Python environment Spyder IDE 5.5.3 to generate the corresponding bar charts. The libraries pandas (The pandas development team 2024) and matplotlib.pyplot (Hunter 2007) were used for data manipulation.

2.7. Mollusk genome completeness assessment

In addition to assessing cephalopod genomes, the cephalopod-specific BUSCO lineage version 2 and coleoid-specific BUSCO lineage were also applied to evaluate the completeness of six selected mollusk genomes, chosen for their relevance and well assembled genomes. The six mollusk species included three bivalves (*Crassostrea gigas*, *Mizuhopecten yessoensis*, *Mytilus coruscus*, and *Anadara kagoshimensis*) and two gastropods (*Chrysomallon squamiferum* and *Achatina immaculata*).

The results after the BUSCO v5.5.0 analyses were then processed in the Python Environment Spyder IDE 5.5.3, with the pandas (The pandas development team 2024) and matplotlib.pyplot (Hunter 2007) libraries used for data manipulation and graph generation, respectively.

2.8. Data availability

Genome and protein FASTA files were retrieved from different databanks. ***O. bimaculoides*** as well as ***D. pealeii*****:** Octopus bimaculoides genome assembly ASM119413v2 -NCBI - NLM (nih.gov) and Doryteuthis pealeii genome assembly UCB_Dpea_1 - NCBI - NLM (nih.gov), respectively. The protein data of ***E. scolopes*** was retrieved from https://metazoa.csb.univie.ac.at/data/v2/. ***N. pompilius*****:** Genomic data for Nautilus pompilius (figshare.com).

The GenBank bioproject and accession numbers of the RNA-seq reads are: ***D.pealeii*****:** BioProject PRJNA641326 (Accession numbers SRR18071786 - SRR18071813). ***E. scolopes*****:** BioProject PRJNA473394 (Accession numbers SRR172521 - SRR172568, SRR172562 was excluded because of downloading problems), BioProject PRJNA257113 (Accession numbers SRR349107, SRR3495048, SRR3495043 - SRR3495046), BioProject PRJNA498343 (Accession numbers SRR8159233 - SRR8159250, SRR8159246 was excluded because of downloading problems). ***O. bimaculoides*****:** BioProject PRJNA846095 (Accession numbers SRR19548824, SRR19548831 - SRR19548834), BioProject PRJNA285380 (Accession numbers SRR2048495 - SRR2048498, SRR2048521 - SRR2048525, SRR2045866, SRR2045870, SRR2047107,

SRR2047109, SRR2047111, SRR2047114, SRR2047116, SRR2047118, SRR2047120, SRR2047122). ***N. pompilius*:** BioProject PRJNA673746 (Accession numbers SRR13132385 - SRR13131288, SRR13131290), BioProject PRJNA614552 (Accession numbers SRR11485675 - SRR11485682, SRR11485684 - SRR11485687), BioProject PRJNA369717 (Accession number SRR5626553), BioProject PRJNA299756 (Accession numbers SRR2857279 - SRR2857280). ***O. vulgaris*:** BioProject PRJEB11391 (Accession numbers ERR1059207 - ERR1059218).

Genomes of other cephalopod species are available at NCBI and GigaDB. ***O. sinensis*:** Octopus sinensis genome assembly ASM634580v1 - NCBI - NLM (nih.gov), ***O. vulgaris*:** Octopus vulgaris genome assembly xcOctVulg1.2 - NCBI - NLM (nih.gov), ***T. rhombus*:** Thysanoteuthis rhombus genome assembly xcThyRhom1.1 - NCBI - NLM (nih.gov), ***S. lessoniana*:** Sepioteuthis lessoniana genome assembly xcSepLess1.1 - NCBI - NLM (nih.gov) and ***S. atlantica*:** Sepiola atlantica genome assembly xcSepAtla11.1 - NCBI - NLM (nih.gov).

Genomes of the other molluscan species are available at NCBI and Dryad. ***C. gigas*:** Magallana gigas genome assembly cgigas_uk_roslin_v1 - NCBI - NLM (nih.gov), ***M. yessoensis*:** Mizuhopecten yessoensis genome assembly ASM211388v2 - NCBI - NLM (nih.gov), ***M. coruscus*:** https://datadryad.org/stash/downloads/file_stream/507455, ***C. squamiferum*:** https://datadryad.org/stash/downloads/file_stream/507454, ***A. immaculata*:** Achatina immaculata genome assembly ASM976088v1 - NCBI - NLM (nih.gov) and ***A. kagoshimensis*:** Anadara kagoshimensis genome assembly ASM2129210v1 - NCBI - NLM (nih.gov).

3. Results

In comparative genomics and evolutionary biology, accurately assessing genome completeness is crucial for analyzing, interpreting biological data, and understanding species evolution. BUSCO has become a widely used tool to evaluate genome completeness by comparing assemblies against a set of highly conserved orthologous genes. The following chapters present the results of analyses using cephalopod- and coleoid-specific sets of high-quality single-copy orthologs, providing insights into the completeness and quality of these genomic data.

3.1. Cephalopod-specific BUSCO lineages

3.1.1. Cephalopod-specific BUSCO lineage version 1

This customized BUSCO lineage was based on protein sequences available from previous genome annotation studies.

OrthoFinder 2.5.5 identified 2,283 single-copy orthologous genes across *D. pealeii*, *E. scolopes*, *O. bimaculoides* and *N. pompilius*. The total numbers of complete (C), single-copy (S), duplicated (D), fragmented (F) and missing (M) genes are detailed in **Tab. 1**. Briefly, initial BUSCO v5.4.6 (Simão *et al.* 2015) analysis yielded the following genome completeness results: ***E. scolopes*** C: 77.0% [S: 76.7%, D: 0.3%], F: 0.0%, M: 23.0%, ***D. pealeii*** C: 87.3% [S: 86.9%, D: 0.4%], F: 0.0%, M: 12.7%, ***O. bimaculoides*** C: 80.6% [S: 80.4%, D: 0.2%], F: 0.0%, M: 19.4% and ***N. pompilius*** C: 68.4% [S: 68.3%, D: 0.1%], F: 0.0%, M: 31.6%.

The initial analysis identified 1,977 missing BUSCOs across all four species, with 1,165 being unique. We examined the bit score values of each missing ortholog to investigate the high number of missing genes. According to BUSCO criteria, a gene is considered a true ortholog if its bit score exceeds 90% of the minimum bit score for the corresponding orthogroup. In *O. bimaculoides* and *N. pompilius* none of the genes identified as missing met the threshold for true orthologs. On the contrary, *D. pealeii* and *E. scolopes* had one and two genes, respectively, that were true orthologs but were nonetheless categorized as missing. Overall, 99.74% (1,162 out of 1,165) of the missing genes were not true orthologs, while only 0.26% (3 out of 1165) were considered missing due to potential prediction artifacts. Notably, one missing gene in *E. scolopes* was linked to annotations on multiple chromosomes, and this gene was not identified as missing in the BUSCO rerun.

After determining that 99.84% (1,163 out of 1,165) of the missing genes were not true orthologs, they were filtered out with the prediction “artifacts”. The final dataset thus contained

1,118 orthologous genes, as 1,165 out of 2,283 were excluded. The completeness of the genome assemblies with the final filtered data was: ***E. scolopes*** C: 99.9% [S: 99.4%, D: 0.5%], F: 0.0%, M: 0.1%, ***D. pealeii*** C: 100.0% [S: 99.4%, D: 0.6%], F: 0.0%, M: 0.0%, ***O. bimaculoides*** C: 99.9% [S: 99.6%, D: 0.3%], F: 0.1%, M: 0.0% and ***N. pompilius*** C: 100.0% [S: 99.8%, D: 0.2%], F: 0.0%, M: 0.0%. As shown in **Tab. 1**, there is only one missing gene remaining in *E. scolopes*. Overall, the number of complete and single-copy orthologs is more consistent across all four species compared to the initial dataset. Additionally, there is an obvious decrease in both missing and duplicated BUSCOs. Initially, *N. pompilius* had the lowest amount of complete and single-copy orthologs and the highest number of missing orthologs, but this trend is not evident in the final dataset after filtering (see **Tab. 1**).

Tab. 1: Gene counts across BUSCO categories for the initial and final versions of the cephalopod-specific BUSCO lineage version 1. Each column in this table contains two values in the format 'initial / final', representing the number of orthologous genes. C...complete, S...single-copy, D...duplicated, F...fragmented, M...missing.

Species	C	S	D	F	M	# BUSCOs
E. scolopes	1759 / 1117	1752 / 1111	7 / 6	0 / 0	524 / 1	2283 / 1118
D. pealeii	1993 / 1118	1984 / 1111	9 / 7	0 / 0	290 / 0	2283 / 1118
O. bimaculoides	1840 / 1117	1835 / 1114	5 / 3	1 / 1	442 / 0	2283 / 1118
N. pompilius	1562 / 1118	1560 / 1116	2 / 2	0 / 0	721 / 0	2283 / 1118

3.1.2. Cephalopod-specific BUSCO lineage version 2

This BUSCO lineage was based on RNA-seq data prepared with the Mikado v2.3.4. pipeline.

For this dataset, OrthoFinder 2.5.5 identified 974 single-copy orthologous genes among *D. pealeii*, *E. scolopes*, *O. bimaculoides* and *N. pompilius*. A preliminary BUSCO v5.4.6 assessment provided the following genome completeness results: ***D. pealeii*** C: 90.5% [S: 89.3%, D: 1.2%], F: 0.0%, M: 9.5%, ***E. scolopes*** C: 77.3% [S: 77.1%, D: 0.2%], F: 0.0%, M: 22.7%, ***O. bimaculoides*** C: 81.5% [S: 81.4%, D: 0.1%], F: 0.0%, M: 18.5% and ***N. pompilius*** C: 65.9% [S: 65.7%, D: 0.2%], F: 0.0%, M: 34.1%. The total counts of complete (C), single-copy (S), duplicated (D), fragmented (F) and missing (M) BUSCOs are shown in **Tab. 2**.

In total, 825 missing BUSCOs were identified, of which 462 are unique. As with the cephalopod-specific BUSCO lineage version 1, we assessed whether the genes identified as missing by BUSCO v5.4.6 were true orthologs. All 462 missing genes did not meet the score cutoff parameter and were therefore excluded from the initial dataset. This resulted in a total number of 512 orthologous genes used to generate the final cephalopod-specific BUSCO lineage version 2.

After excluding 462 missing genes, BUSCO v5.4.6 results were as follows: ***E. scolopes*** C. 100.0% [S: 99.8%, D: 0.2%], F: 0.0%, M: 0.0%, ***D. pealeii*** C: 100.0% [S: 98.0%, D: 2.0%], F:

0.0%, M: 0.0%, ***O. bimaculoides*** C: 100.0% [S: 99.8%, D: 0.2%], F: 0.0%, M: 0.0% and ***N. pompilius*** C: 100.0% [S: 99.6%, D: 0.4%], F: 0.0%, M: 0.0%.

Comparing the initial and final datasets revealed a decrease in missing genes. The number of duplicated BUSCOs decreased for *E. scolopes* and *D. pealeii*, while it remained unchanged for the other species. Initially, *N. pompilius* had the highest number of missing BUSCOs (332 genes), whereas *D. pealeii* had the fewest missing BUSCOs (92 genes). Additionally, in this BUSCO lineage version, *N. pompilius* initially had the lowest number of complete and single-copy orthologs. Overall, none of the four species had fragmented or missing BUSCO genes. All species exhibited the same number of complete BUSCOs, with minor differences noted in single-copy and duplicated BUSCOs (see **Tab. 2**).

Tab. 2: Gene counts across BUSCO categories for the initial and final versions of the cephalopod-specific BUSCO lineage version 2. Each column in this table contains two values in the format 'initial / final', representing the number of orthologous genes. C...complete, S...single-copy, D...duplicated, F...fragmented, M...missing.

Species	C	S	D	F	M	# BUSCOs
E. scolopes	753 / 512	751 / 511	2 / 1	0 / 0	221 / 0	974 / 512
D. pealeii	882 / 512	870 / 502	12 / 10	0 / 0	92 / 0	974 / 512
O. bimaculoides	794 / 512	793 / 511	1 / 1	0 / 0	180 / 0	974 / 512
N. pompilius	642 / 512	640 / 510	2 / 2	0 / 0	332 / 0	974 / 512

3.1.3. Comparison of the cephalopod-specific sets of orthologs

The final cephalopod-specific BUSCO lineage version 1 included 1,118 cephalopod-specific orthologous genes for evaluating the completeness of cephalopod genome assemblies (see **Tab. 1**), while the cephalopod-specific BUSCO lineage version 2 provided 512 cephalopod-specific orthologs (see **Tab. 2**). For *E. scolopes* 94.28% of the orthologs identified by version 1 are located on the largest 1-46 scaffolds, and 5.10% on other shorter scaffolds. In comparison, orthologs from version 2 in *E. scolopes* are found to be exclusively on chromosome-scale scaffolds 1-45, with a higher proportion of 97.08% on these scaffolds and 2.73%, on other scaffolds (see **Tab. 3**).

In *D. pealeii*, the percentage of orthologs on chromosome-scale scaffolds is higher for version 1 (99.02%) compared to version 2 (97.27%). Conversely, the proportion of orthologs on non-chromosome-scale scaffolds is lower with version 1 (0.36%) compared to version 2 (0.78%) (see **Tab. 3**).

For *O. bimaculoides* and *N. pompilius*, version 2 exhibits greater coverage of orthologs on chromosome-scale scaffolds, with 99.61% for *O. bimaculoides* and 99.41% for *N. pompilius*. In comparison, version 1 shows lower coverage with 98.48% for *O. bimaculoides* and 99.38% for *N. pompilius*. The percentage of orthologs on non-chromosome-scale scaffolds was also

lower, with 0.20% for *O. bimaculoides* version 2 versus 1.16% with version 1, and 0.20% for *N. pompilius* versus 0.45% with version 1 (see **Tab. 3**). Overall, version 2 of the cephalopod-specific BUSCO lineage demonstrated higher coverage of orthologous genes on chromosome-scale scaffolds in three out of four species compared to version 1.

Tab. 3: Comparisons of genomic and cephalopod-specific orthologous gene characteristics between cephalopod-specific BUSCO lineage version 1 (v1) and version 2 (v2). v1 included 1,118 single-copy orthologs, while v2 contained 512 single-copy orthologs. For columns with two values, the format is 'v1 / v2', representing the number of orthologs on chromosome-scale scaffolds or remaining scaffolds. Chr...chromosomes, OGs...orthologs, w/...with.

Species	# Sequences	# Chr.	# Scaffolds	# OGs on Chr.	# OGs on Scaffolds	# Chr. w/ OGs
E. scolopes	1 735	**46**	1 689	**1054 / 497**	57 / 14	**1-46**
D. pealeii	2 138	**46**	2 092	**1107 / 498**	4 / 4	**1-46**
O. bimaculoides	145 327	**30**	145 296	**1101 / 510**	13 / 1	**1-30**
N. pompilius	3 130	**26**	3 104	**1111 / 509**	5 / 1	**1-26**

The following observations emerged from analyzing the length of orthologs across both BUSCO lineage versions. Cephalopod-specific BUSCO lineage version 1 contained more than twice as many single-copy orthologs compared to version 2, and the mean length of orthogroups was also greater. Specifically, orthogroups in version 1 ranged from 49.5 to 137.7 bp, while those in version 2 ranged from 47.4 to 59.8 bp. Additionally, the median length of orthogroups in version 1 was 79 bp, which was higher than the median length of 60 bp observed in version 2 (see **Fig. 6**).

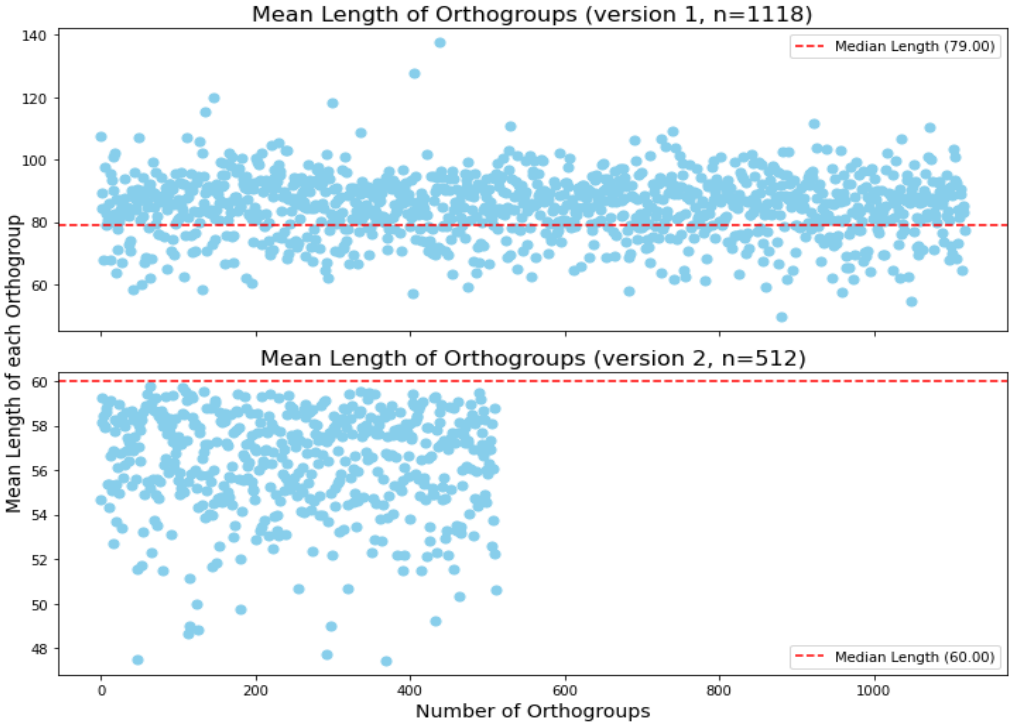

Fig. 6: Comparisons of the mean lengths of each orthogroup from the cephalopod-specific BUSCO lineage version 1 (upper plot) and version 2 (lower plot). The scatter plots illustrate the distribution of mean lengths for each orthogroup. The red dashed line in each plot indicates the median lengths.

3.2. Coleoid-specific BUSCO lineage

To investigate variations in orthologous gene content across different cephalopod species, we replaced the initial outgroup *N. pompilius* with the octopod species *O. vulgaris*, thereby generating a coleoid-specific BUSCO lineage. OrthoFinder 2.5.5 was utilized to identify 1,230 single-copy orthologs across the species *E. scolopes*, *D. pealeii*, *O. bimaculoides* and *O. vulgaris*. A preliminary BUSCO v5.5.0 assessment provided the following genome completeness results: ***E. scolopes*** C:73.3% [S:72.7%, D:0.6%], F:0.0%, M:26.7%, ***D. pealeii*** C:83.5% [S:82.4%, D:1.1%], F:0.0%, M:16.5%, ***O. bimaculoides*** C:83.4% [S:83.3%, D:0.1%], F:0.0%, M:16.6% and ***O. vulgaris*** C:86.9% [S:86.3%, D:0.6%], F:0.0%, M:13.1%. Total counts of complete (C), single-copy (S), duplicated (D), fragmented (F) and missing (M) genes are provided in **Tab. 4**.

In total, 898 orthologs were identified as missing, of which 484 are unique. After excluding the unique missing genes, 746 coloeid-specific orthologs were used to generate the final dataset. The completeness of this dataset was evaluated and returned the following results: ***E. scolopes*** C:100.0% [S:99.2%, D:0.8%], F:0.0%, M:0.0%, ***D. pealeii*** C:100.0% [S:98.5%, D:1.5%], F:0.0%, M:0.0%, ***O. bimaculoides*** C:100.0% [S:99.9%, D:0.1%], F:0.0%, M:0.0% and ***O. vulgaris*** C:100.0% [S:99.1%, D:0.9%], F:0.0%, M:0.0%.

Comparisons of the initial and final dataset revealed a major reduction in the number of missing genes and complete genes, which was expected given the filtering of missing genes in the initial dataset. The initial version showed that *E. scolopes* had the highest number of missing genes and the lowest counts of complete and single-copy orthologs. In the final dataset, the counts of complete and single-copy genes were largely consistent across all species, with only minor differences. Additionally, both versions of the dataset showed no fragmented orthologs. There was also a decrease in duplicated orthologs in both *E. scolopes* and *D. pealeii* (see. **Tab. 4**).

Tab. 4: Gene counts across BUSCO categories for the initial and final versions of the coleoid-specific BUSCO lineage. Each column in this table contains two values in the format 'initial / final', representing the number of orthologous genes. C...complete. S...single-copy, D...duplicated, F...fragmented, M...missing.

Species	C	S	D	F	M	# BUSCOs
E. scolopes	901 / 746	894 / 740	70 / 6	0 / 0	329 / 0	1230 / 746
D. pealeii	1026 / 746	1013 / 735	13 / 11	0 / 0	204 / 0	1230 / 746
O. bimaculoides	1026 / 746	1025 / 745	1 / 1	0 / 0	204 / 0	1230 / 746
O. vulgaris	1069 / 746	1062 / 739	7 / 7	0 / 0	161 / 0	1230 / 746

3.2.1. Cephalopod-specific *versus* coleoid-specific BUSCO lineage

Overall, no fragmented and missing genes were present in any of the datasets, except for a single fragmented gene in *O. bimaculoides* and a single missing gene in *E. scolopes* with the

cephalopod-specific BUSCO lineage version 1 (see **Tab. 1**). The percentage of duplicated orthologs was 0.40% in the cephalopod-specific BUSCO lineage version 1, 0.68% in version 2, and 0.84% in the coleoid-specific BUSCO lineage. Consequently, the cephalopod-specific BUSCO lineage version 1 had the highest percentage of single-copy orthologs at 99.55%, followed by version 2 at 99.32%, and the coleoid-specific BUSCO lineage at 99.16%.

When comparing the two datasets based on RNA-seq data, the coleoid-specific BUSCO lineage contained a higher total number of orthologous genes (746 in total) compared to the cephalopod-specific BUSCO lineage version 2 (512 in total) (see **Tab. 2 & 4**). However, the cephalopod-specific BUSCO lineage version 2 exhibited a higher percentage of single-copy and duplicated orthologous genes.

3.3. Assessing cephalopod genome completeness

During the generation of the cephalopod-specific core set of orthologous genes, the alignment software Miniprot v0.12 (Li 2023) was newly available within BUSCO v5.5.0. The cephalopod-specific BUSCO lineage version 2 was processed with Miniprot v0.12, adding a better gene search compared to the initial annotation tool MetaEuk (Levy Karin *et al.* 2020). Furthermore, newly released chromosome-scale genome assemblies were analyzed using both MetaEuk and Miniprot v0.12.

As shown in **Tab. 5**, both datasets produced identical orthologous genes, with 512 cephalopod-specific orthologs. However, results with Miniprot v0.12 showed a higher number of single-copy orthologs and fewer duplications and missing orthologs compared to MetaEuk. Additionally, for a better comparison of dataset quality, results using the metazoan BUSCO lineage (metazoa_odb10) were also included in **Tab. 5**. Despite including more single-copy orthologs in the metazoan dataset, its performance was poorer compared to the cephalopod-specific dataset, demonstrating a higher rate of missing, duplicated and fragmented orthologs (see **Tab. 5**). When examining the distribution of orthologous genes across chromosome-scale scaffolds, Miniprot v0.12 generally identified more orthologs on these scaffolds than MetaEuk, except for *E. scolopes*, *O. bimaculoides* and *N. pompilius*. For *E. scolopes* and *O. bimaculoides* Miniprot v0.12 identified only one additional gene on chromosome-scale scaffolds compared to MetaEuk, while *N. pompilius* had five more genes. MetaEuk also provided fewer or an equal number of orthologs on other scaffolds (see **Tab. 6**). The distribution of orthologs across chromosome-scale scaffolds was similar for both software. *E. scolopes* had orthologous genes distributed across the 45 largest chromosome-scale scaffolds, with none present on the 46th scaffold. Due to incomplete genome assembly for *S. atlantica* and *S. lessoniana*, chromosome-scale scaffolds 42-46 and 45-46 were missing, respectively, and as

a result, no orthologs were present on these scaffolds. For all other species, orthologs were present on every chromosome-scale scaffold (see **Tab. 6**).

Tab. 5: Comparison of gene counts across BUSCO categories between MetaEuk and Miniprot v0.12 with the cephalopod-specific dataset (n=512), and metazoa_odb10 (n=954) with Miniprot v0.12. Each column in this table contains three values in the format 'MetaEuk / Miniprot / metazoa_odb10', representing the number of orthologous genes identified for each method. Species used to generate the cephalopod-specific BUSCO lineage version 2 are highlighted in green. C...complete, D...duplicated, F...fragmented, M...missing.

Species	C	S	D	F	M
E. scolopes	512 / 512 / 908	511 / 512 / 875	1 / 0 / 33	0 / 0 / 4	0 / 0 / 42
D. pealeii	512 / 512 / 907	502 / 511 / 825	10 / 1 / 82	0 / 0 / 12	0 / 0 / 35
S. atlantica	483 / 499 / 905	479 / 498 / 895	4 / 1 / 10	0 / 0 / 8	29 / 13 / 41
S. lessoniana	462 / 472 / 890	459 / 471 / 886	3 / 1 / 4	0 / 0 / 30	50 / 40 / 34
T. rhombus	485 / 501 / 934	482 / 500 / 929	3 / 1 / 5	0 / 0 / 12	27 / 11 / 8
O. bimaculoides	512 / 510 / 940	511 / 510 / 909	1 / 0 / 31	0 / 0 / 2	0 / 2 / 12
O. vulgaris	479 / 497 / 909	477 / 496 / 903	2 / 1 / 6	0 / 0 / 12	33 / 15 / 33
O. sinensis	491 / 506 / 942	490 / 506 / 906	1 / 0 / 36	0 / 0 / 5	21 / 6 / 7
N. pompilius	512 / 504 / 938	510 / 503 / 896	2 / 1 / 42	0 / 0 / 3	0 / 8 / 13

Tab. 6: Genomic characteristics and cephalopod-specific orthologous gene characteristics between MetaEuk and Miniprot v0.12. The format of columns with two values is as follows: 'MetaEuk / Miniprot', representing the number of orthologs for each software. Highlighted in green are the species used to generate the cephalopod-specific BUSCO lineage version 2. Chr...chromosomes, OGs...orthologs, w/...with.

Species	# Sequences	# Chr.	# Scaffolds	# OGs on Chr.	# OGs on Scaffolds	# Chr. w/ OGs
E. scolopes	1 735	**46**	1 689	**497 / 496**	14 / 16	**1-45**
D. pealeii	2 138	**46**	2 092	**498 / 512**	4 / 5	**1-46**
S. atlantica	1 380	**41**	1 338	**476 / 497**	3 / 3	**1-41**
S. lessoniana	6 820	**44**	6 775	**408 / 419**	51 / 54	**1-44**
T. rhombus	621	**46**	574	**480 / 497**	2 / 4	**1-46**
O. bimaculoides	145 327	**30**	145 296	**510 / 509**	1 / 1	**1-30**
O. vulgaris	226	**30**	196	**477 / 497**	0 / 0	**1-30**
O. sinensis	13 516	**30**	13 485	**480 / 495**	10 / 11	**1-30**
N. pompilius	3 130	**26**	3 104	**509 / 504**	1 / 1	**1-26**

In addition to evaluating the cephalopod-specific BUSCO dataset, the genome assemblies of above-mentioned cephalopods were also analyzed using other available BUSCO lineages to compare their effectiveness in assessing genome completeness. As illustrated in **Fig. 7**, the cephalopod-specific BUSCO lineage version 2 provided a more accurate measurement of genome completeness of cephalopod assemblies compared to the eukaryotic (eukaryota_odb10), metazoan (metazoa_odb10) and molluscan (mollusca_odb10) datasets, with the metazoan dataset being more suitable than the eukaryotic or even the molluscan one. Using the cephalopod-specific lineage, almost no duplications and no fragmentations at all were detected in the selected genomes, with only a few missing genes identified. In contrast, the eukaryotic, metazoan and molluscan datasets revealed both duplicated and fragmented

orthologous genes. The molluscan BUSCO lineage identified the highest number of missing orthologs across all species compared to the other three datasets (see **Fig. 7**).

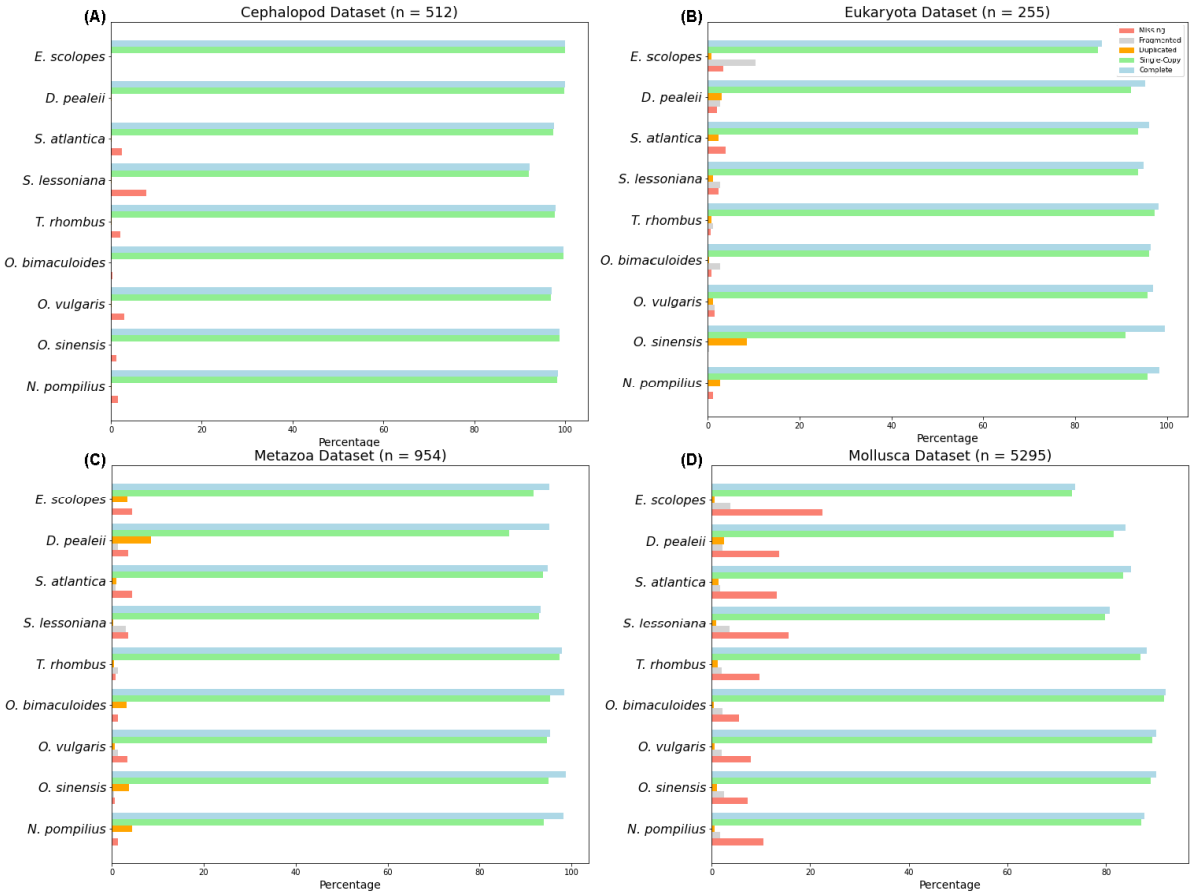

Fig. 7: Comparison between the **(A)** cephalopod-specific, **(B)** eukaryotic, **(C)** metazoan and **(D)** molluscan BUSCO lineage.

To gain deeper insight into the total numbers of orthologs assessed with the core set of cephalopod-specific orthologous genes, the completeness of the above-mentioned cephalopod genome assemblies was analyzed and is illustrated in **Fig. 8**. No fragmented genes were detected in any of these species. Although there are missing genes in some species used to generate this core set of genes – specifically two missing genes in *O. bimaculoides* and eight in *N. pompilius* – the number of missing genes is lower compared to other species. The highest number of missing orthologs was observed in *S. lessoniana*. Among the species not used to generate the cephalopod-specific dataset, *Octopus sinensis* had the lowest number of missing genes, with six genes identified. In contrast, *S. atlantica*, *T. rhombus* and *O. vulgaris* had 13, 11 and 15 missing orthologs, respectively. As noted, no fragmentation was found, indicating that the remaining genes across all species are either complete, single-copy or duplicated. Six species, including two used to generate the core gene set, each had one duplicated ortholog, with the rest being single-copy orthologs (see **Fig. 8**).

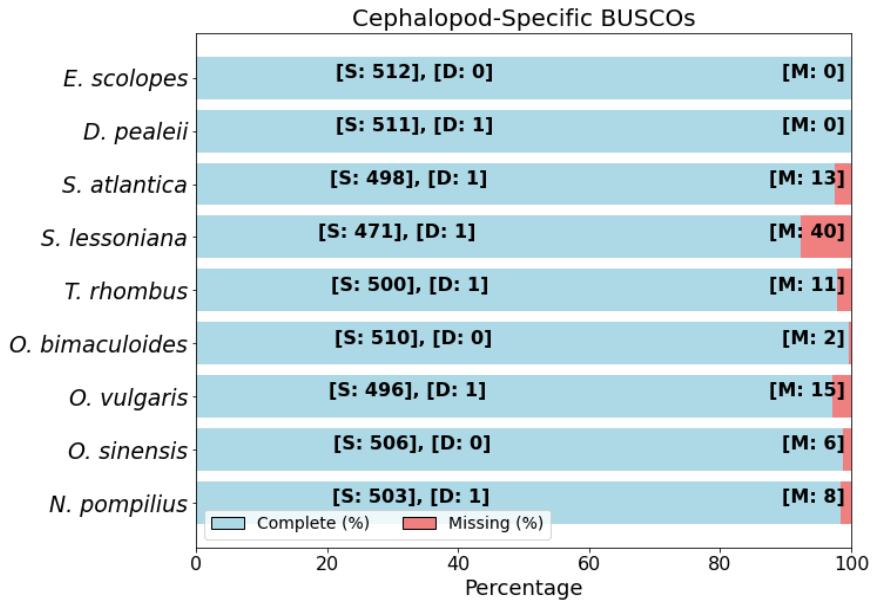

Fig. 8: Cephalopod genome completeness using the core set of cephalopod-specific orthologous genes with BUSCO v5.5.0 and Miniprot v0.12.

In addition to comparing the completeness assessment results between different BUSCO lineages, we analyzed the number of orthologs shared between these lineages. For this analysis, we included the eukaryotic, metazoan and molluscan datasets alongside the cephalopod- and coleoid-specific datasets. The cephalopod-specific BUSCO lineage version 2 was based on *E. scolopes*, *D. pealeii*, *O. bimaculoides* and *N. pompilius*, while the coleoid-specific BUSCO lineage substituted *N. pompilius* with *O. vulgaris*. **Figure 9**, presented as a Venn diagram, illustrates the shared and unique orthologous genes across all species and datasets.

The eukaryotic dataset contained 255 single-copy orthologs, the metazoan dataset contained 954, and the molluscan dataset included 5,295. The cephalopod-specific dataset utilized 512 single-copy orthologs, while the coleoid-specific dataset included 746.

In the cephalopod-specific part of the Venn diagram, 117 orthologs were unique to the cephalopod dataset. This dataset shared 13 orthologs with the metazoan, 256 with the molluscan, and only three with the eukaryotic dataset. Notably, only 25 single-copy orthologs were present in all four BUSCO lineages (see **Fig. 9A**).

The ellipse representing the coleoid-specific part of the Venn diagram shows that 212 orthologous genes were unique to this dataset. The coleoid dataset shared four orthologs with the eukaryotic, 21 with the metazoan, and 361 with the molluscan BUSCO dataset. A total of 29 orthologous genes were shared among all BUSCO lineages (see **Fig. 9B**).

Both customized BUSCO datasets, specific for cephalopods or coleoids, revealed that the coleoid-specific dataset contained more orthologs (746 in total) than the cephalopod-specific dataset (512 in total). Consequently, there were more unique coleoid-specific genes (212 in total) than cephalopod-specific ones (117 in total). Additionally, the coleoid-specific BUSCO lineage shared more genes with the other lineages compared to the cephalopod-specific lineage (see **Fig. 9**).

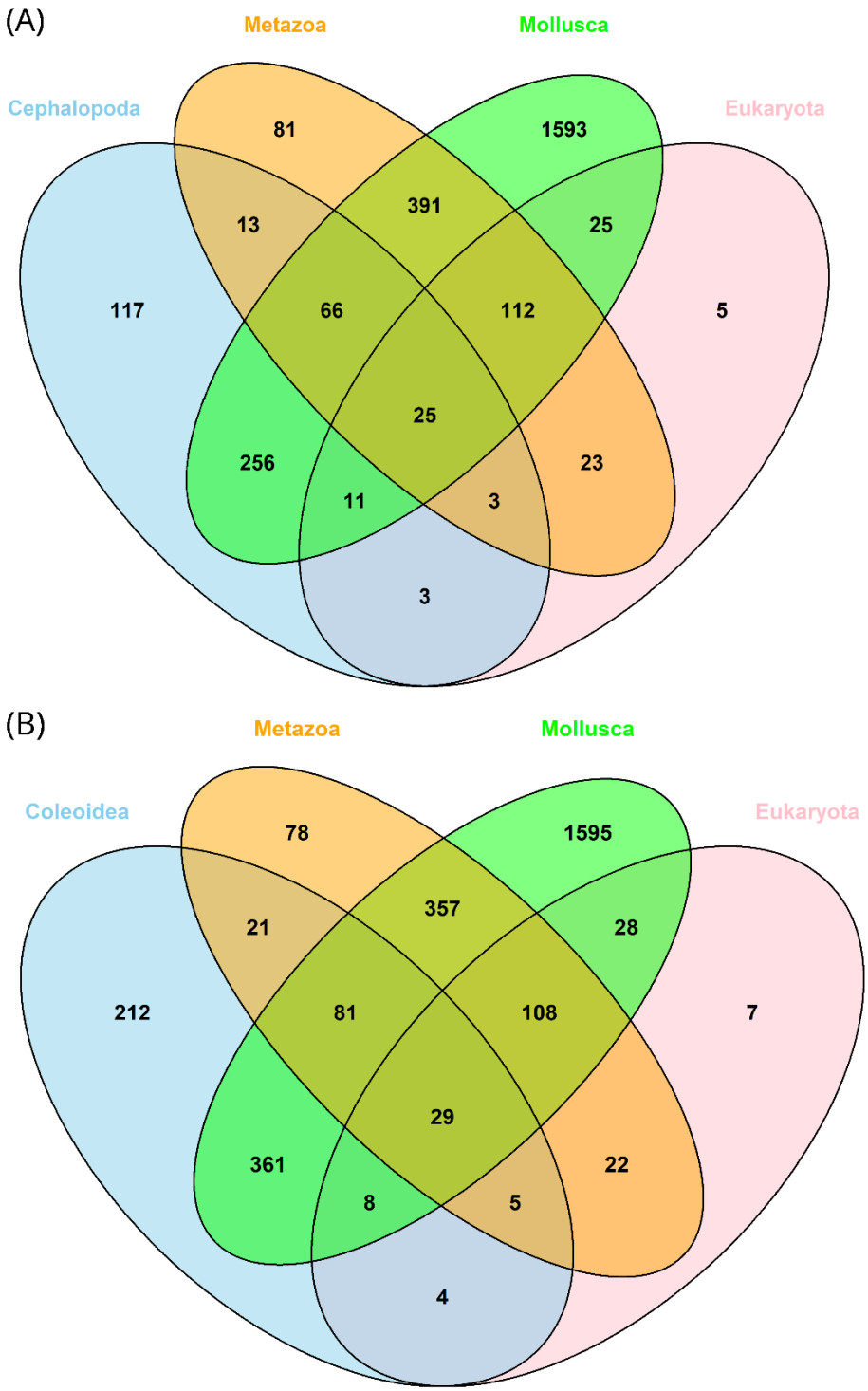

Fig. 9: Venn diagram of the **(A)** cephalopod-specific and **(B)** coleoid-specific BUSCO lineages. The numbers inside represent the shared or unique gene counts across these lineages.

Oxford dot plots derived from the core sets of cephalopod-specific and coleoid-specific single-copy orthologs provided valuable insights into the positional relationships of orthologous genes among homologous chromosomes, also known as synteny. For instance, as illustrated in **Fig. 10**, a diagonal line was observed when comparing *E. scolopes* and *D. pealeii*, indicating conserved synteny between the two species. In general, orthologs that align along a straight diagonal line follow same synteny in both species, whereas counter-diagonal lines represent translocations within the same chromosome. Completely relocation of genes, were recognizable when a set of gene in one species occurred in two or more chromosomes in the other species.

In this comparison, 492 orthologous genes were shared between *E. scolopes* and *D. pealeii*, distributed across all chromosomes, with the remaining 20 orthologs located on other scaffolds (see **Fig. 10A**). Notably, several orthologs on chromosome 2 of the bobtail squid aligned with genes on chromosome 24 and chromosome 40 in *D. pealeii*, suggesting potential chromosomal fusion rearrangements.

A slightly different pattern emerged when using the coleoid-specific gene set. A total of 710 orthologs were shared across all chromosomes, with 36 orthologs remained on other scaffolds. While no gene relocations were identified, several translocations were observed too (see **Fig. 10B**). Additionally, the same chromosomes did not align with the same homologous chromosomes between the two datasets, highlighting further differences in gene arrangements between the species (see **Fig. 10**).

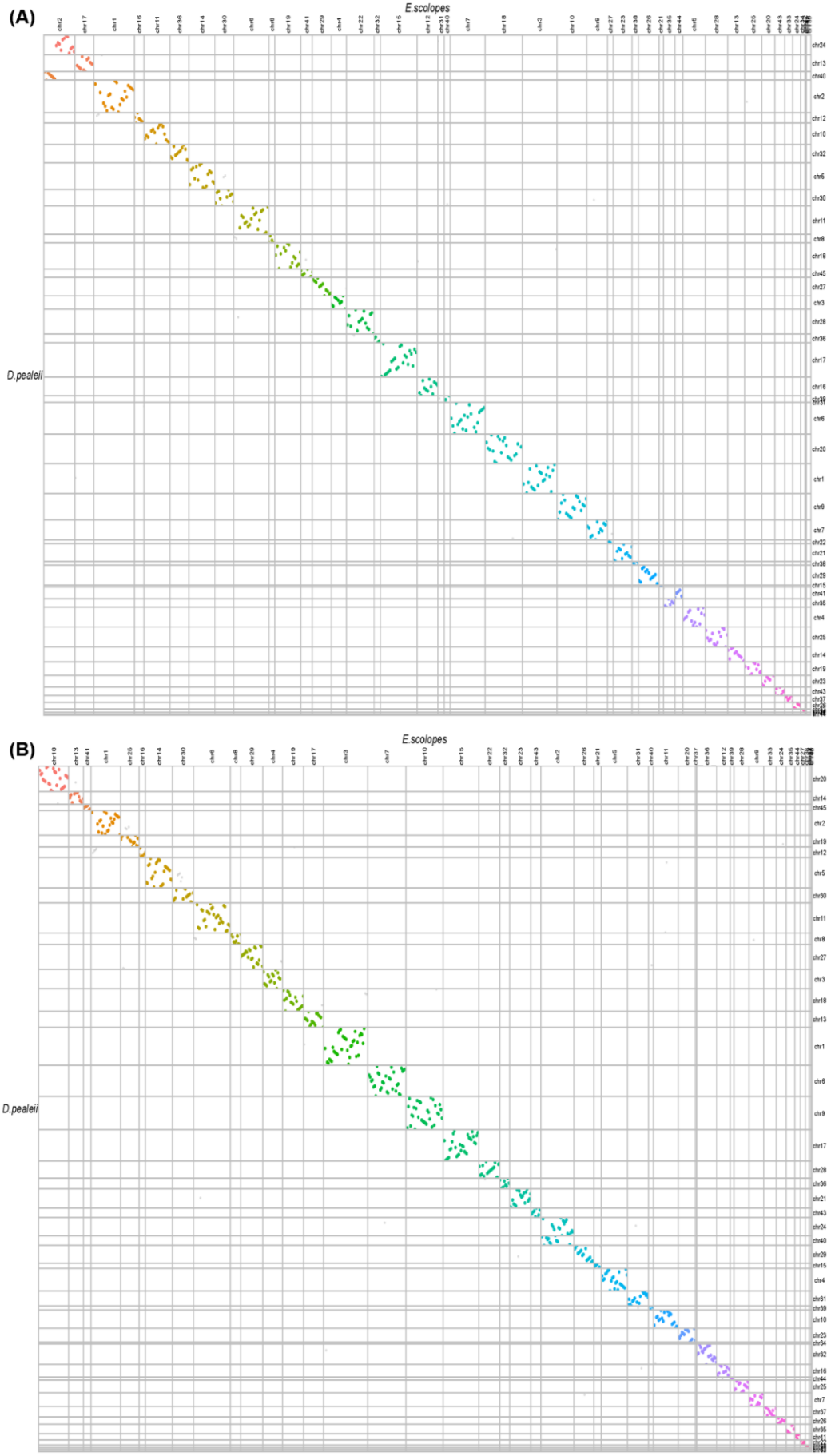

Fig. 10: Synteny comparison between *E. scolopes* (5.1 Gb) and *D. pealeii* (4.6 Gb). **(A)** shows the comparison based on the cephalopod-specific dataset (n=512) and **(B)** based on the coleoid-specific dataset (n=746).

The comparison between *E. scolopes* and *S. atlantica* revealed markedly different results. Using the cephalopod-specific gene set, 484 orthologs were shared across all 46 chromosomes in *E. scolopes* and across 41 chromosomes in *S. atlantica*, with 28 orthologs remaining on other scaffolds or missing (see **Fig. 11A**). The coleoid-specific gene set resulted in 690 shared genes across the chromosomes, with 56 remained on scaffolds or missing (see **Fig. 11B**).

Both gene sets exhibited more gene relocations and translocations in the comparison between *E. scolopes* and *S. atlantica* than in the comparison between *E. scolopes* and *D. pealeii*. For example, orthologs, from the cephalopod-specific gene set on chromosome 10 in *S. atlantica* corresponded to genes on chromosomes 6 and 9 in *E. scolopes*. Additionally, chromosome 20 in *S. atlantica* had its counterpart in chromosome 30 in *E. scolopes* (see **Fig. 11A**).

When using the coleoid-specific gene set, chromosome 6 in *E. scolopes* corresponded to chromosomes 10 and 29 in *S. atlantica*. Conversely, genes on chromosome 2 in *S. atlantica* mapped to chromosomes 10, 17, and 31 in *E. scolopes*. Further potential relocations included genes on chromosome 4 in *S. atlantica*, which had homologs on chromosome 14 and 36 in *E. scolopes*. Additionally, orthologs on chromosomes 18, 29 and 39 in *E. scolopes* aligned with chromosome 5 in *S. atlantica* (see **Fig. 11B**).

Notably, *S. atlantica* had only 41 assembled chromosomes, which may had contributed to the observed relocation and translocation patterns. Overall, the comparison between *E. scolopes* and *S. atlantica* exhibited more translocations than the comparison between *E. scolopes* and *D. pealeii* (see **Fig. 11**).

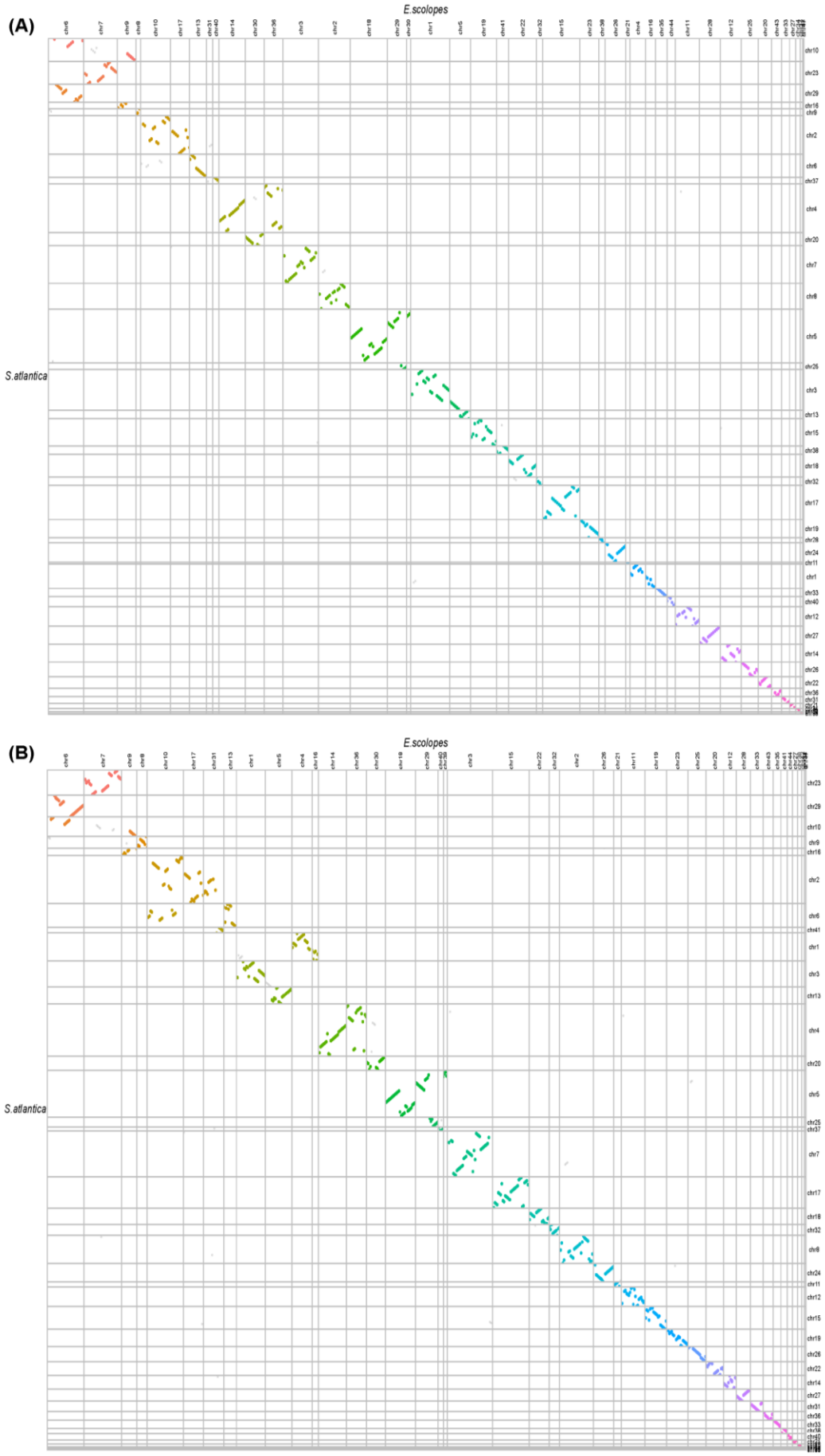

Fig. 11: Synteny comparison between *E. scolopes* (5.1 Gb) and *S. atlantica* (5.6 Gb). **(A)** shows the comparison based on the cephalopod-specific dataset (n=512) and **(B)** based on the coleoid-specific dataset (n=746).

A comparison between *E. scolopes* and *S. lessoniana* using the cephalopod-specific gene set revealed 407 shared genes distributed across 46 chromosomes in *E. scolopes* and 44 chromosomes in *S. lessoniana*. The remaining 105 orthologs were either located on other scaffolds or were missing (see **Fig. 12A**). In contrast, the coleoid-specific gene set identified 566 shared orthologs across these chromosomes, with 180 orthologs found on other scaffolds or missing (see **Fig. 12B**). Unlike the comparison between *E. scolopes* and *S. atlantica*, no significant relocations of orthologs across different chromosomes were observed between *E. scolopes* and *S. lessoniana*. However, gene translocations within the same chromosomes were identified in both gene sets (see **Fig. 12**).

Similarly, the comparison between *E. scolopes* and *T. rhombus* did not show significant anomalies. Using the cephalopod-specific gene set, 483 shared orthologs were identified, with 29 orthologs either remaining on other scaffolds or missing (see **Fig. 13A**). The coleoid-specific gene set provided a total of 566 shared orthologs, with 180 orthologs remaining on other scaffolds or missing (see **Fig. 13B**). As expected for squid species, *T. rhombus* had 46 assembled chromosomes.

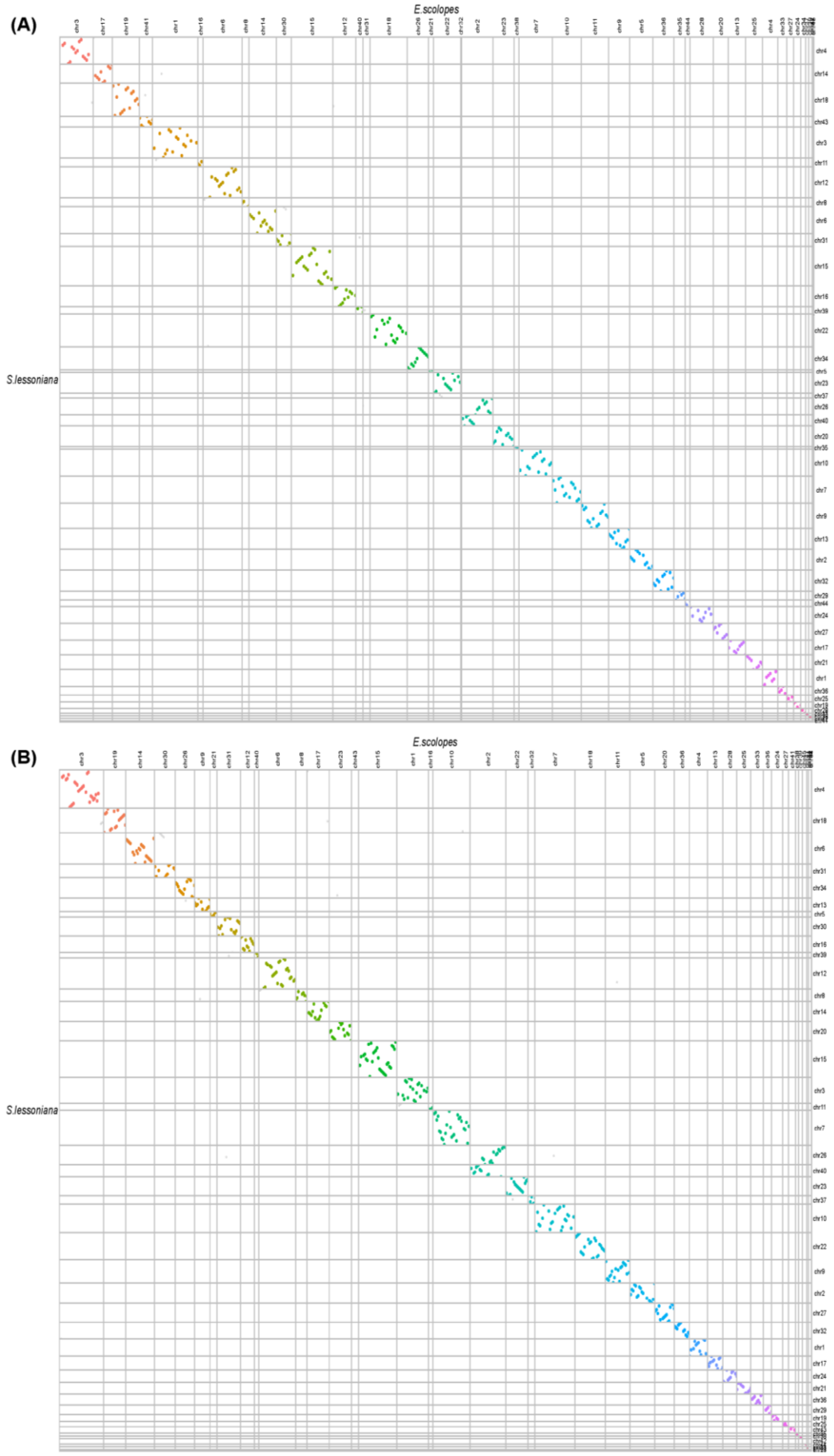

Fig. 12: Synteny comparison between *E. scolopes* (5.1 Gb) and *S. lessoniana* (5.1 Gb). **(A)** shows the comparison based on the cephalopod-specific dataset (n=512) and **(B)** based on the coleoid-specific dataset (n=746).

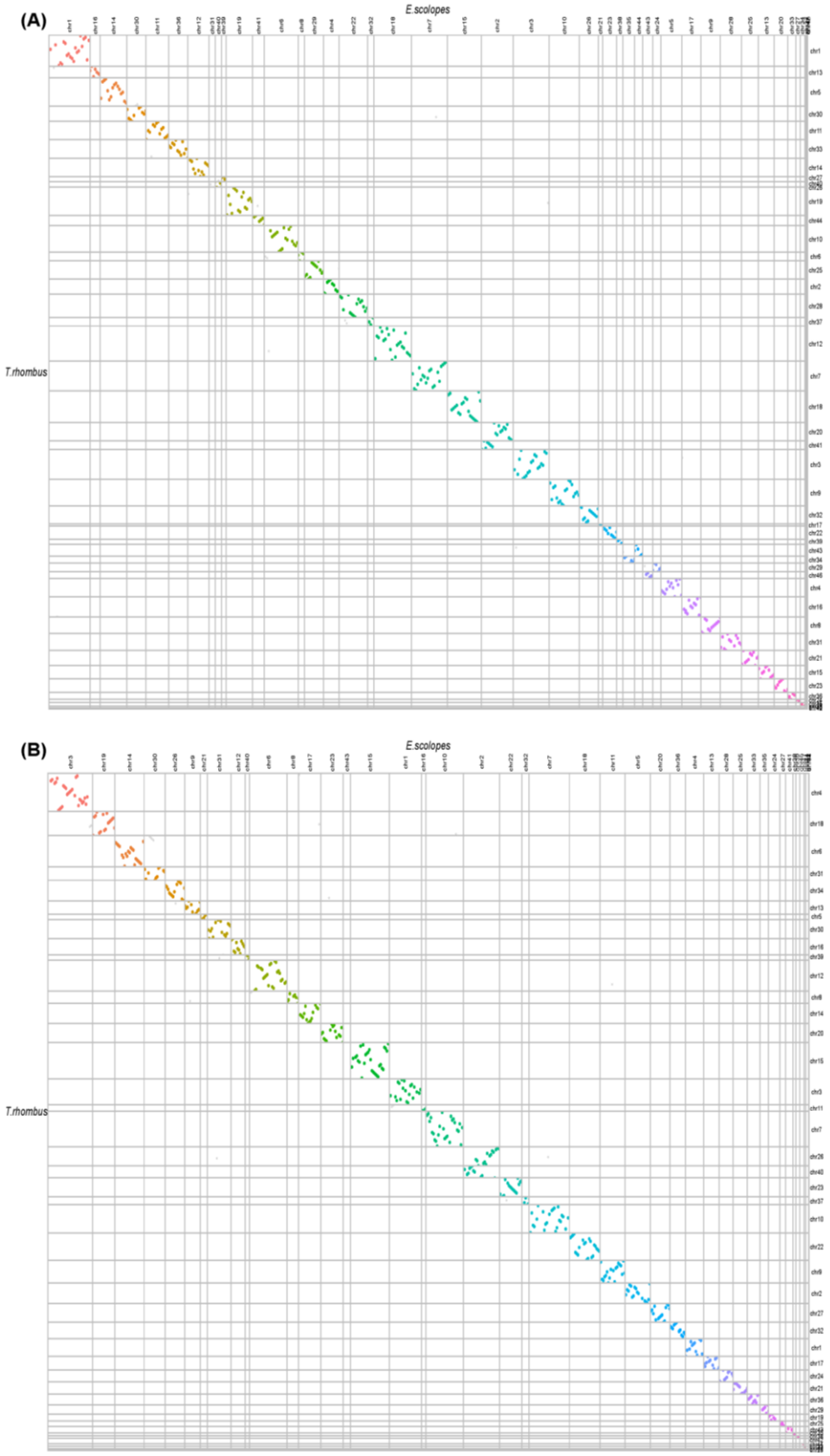

Fig. 13: Synteny comparison between *E. scolopes* (5.1 Gb) and *T. rhombus* (2.7 Gb). **(A)** shows the comparison based on the cephalopod-specific dataset (n=512) and **(B)** based on the coleoid-specific dataset (n=746).

3.3.1. Gene ontology

A Gene Ontology (GO) enrichment analysis (Thomas *et al.* 2022) was conducted using the core set of cephalopod- and coleoid-specific orthologs to identify protein classes, their biological processes, molecular functions and cellular components.

Initially, the specific sets of genes were blasted to identify homologous genes in known species, which were then used for the GO term analysis. For blasting, 117 unique cephalopod-specific genes (see **Fig. 9A**) and 212 unique coleoid-specific genes (see **Fig. 9B**) were utilized. The blast analysis of the cephalopod-specific orthologs revealed 97 out of 117 unique genes, while the coleoid-specific orthologs yielded 148 out of 212 unique genes, resulting in 82.91% uniqueness in cephalopod-specific genes and 69.81% uniqueness in coleoid-specific orthologs.

Following the GO term analysis based on the cephalopod-specific orthologous genes, 23 genes were mapped, while 74 remained unmapped. The mapped orthologs were further analyzed for their protein classes, biological processes, molecular functions and cellular components. In terms of protein classes, twelve distinct classes were observed. Five out of twelve classified genes belonged to the “protein modifying enzyme” class. Cytoskeletal proteins, metabolite interconversion enzymes, and transporters each had two assigned orthologs. The remaining protein classes contained only one ortholog each. Four genes could not be specifically classified (see **Fig. 14**). Regarding the biological processes, eight distinct processes were found. The majority of orthologs were associated with cellular processes (nine genes in total), followed by biological regulations (six in total), metabolic processes (five in total), and responses to stimuli (four in total). Each of the remaining categories contained only one ortholog. Ten orthologs could not be assigned to any biological process. For molecular functions, eight were identified as playing a role in binding activities, and five as playing a role in catalytic activities. The majority of orthologs (ten genes in total) could not be classified into any specific molecular function. The remaining categories included one or two orthologs each. In terms of cellular components, the majority of orthologs were assigned to the cellular anatomical entity components. Additionally, five orthologs were categorized as part of protein-containing-complexes, while five could not be assigned to any cellular components at all (see **Fig. 14**).

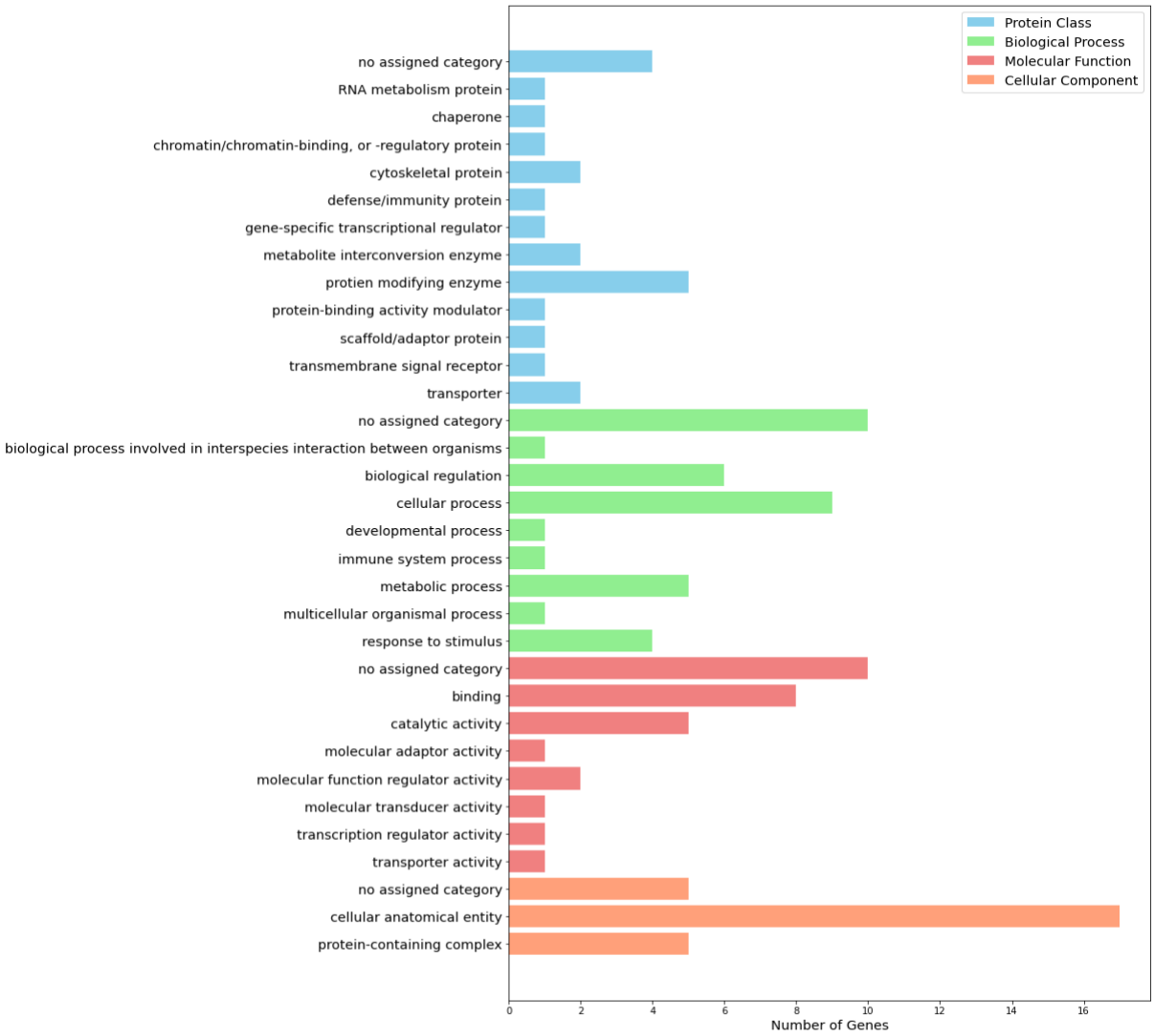

Fig. 14: Gene ontology (GO) enrichment analysis results based on the core set of cephalopod-specific single-copy orthologous genes.

The GO term analysis results for coleoid-specific orthologous genes showed patterns like those of the cephalopod-specific genes. A total of 47 genes were mapped, while 101 remained unmapped. Among the mapped genes, 13 distinct protein classes were identified. Seven orthologs were classified as protein-modifying enzymes, followed by six classified as metabolite interconversion enzymes. Four orthologs each were assigned to the protein classes “protein-binding activity modulator” and “transporters”. The remaining protein classes contained up to three different orthologs each. However, most mapped genes (ten genes in total) could not be classified into any of these protein classes (see **Fig. 15**). Regarding biological processes, twelve distinct processes were observed. Most genes were involved in cellular processes, like the cephalopod-specific genes, followed by metabolic processes (17 in total). Many orthologs were assigned to biological regulation processes (ten in total), while the remaining categories contained up to four genes each. Additionally, a high number of genes (18 genes in total) were not assigned to any biological process. In terms of molecular functions,

seven distinct functions were identified. However, most genes (22 in total) could not be assigned to any specific function. As with the cephalopod-specific orthologs, most coleoid-specific genes exhibited binding (twelve genes in total) and catalytic activities (eleven in total), with the remaining categories containing up to three genes each. For cellular components, most orthologs (30 genes in total) were assigned to the cellular anatomical entity component. Additionally, ten were classified as part of the protein-containing complex, while 14 were not assigned to any component at all (see **Fig. 15**).

Overall, there were only slight differences between the cephalopod-specific and coleoid-specific sets of orthologs concerning protein classes, biological processes, molecular functions, and cellular components. Furthermore, 23.71% of cephalopod-specific genes and 31.76% of coleoid-specific genes could be mapped. Additionally, 47 out of 97 cephalopod-specific genes were present in the coleoid-specific gene set, consisting of 148 genes.

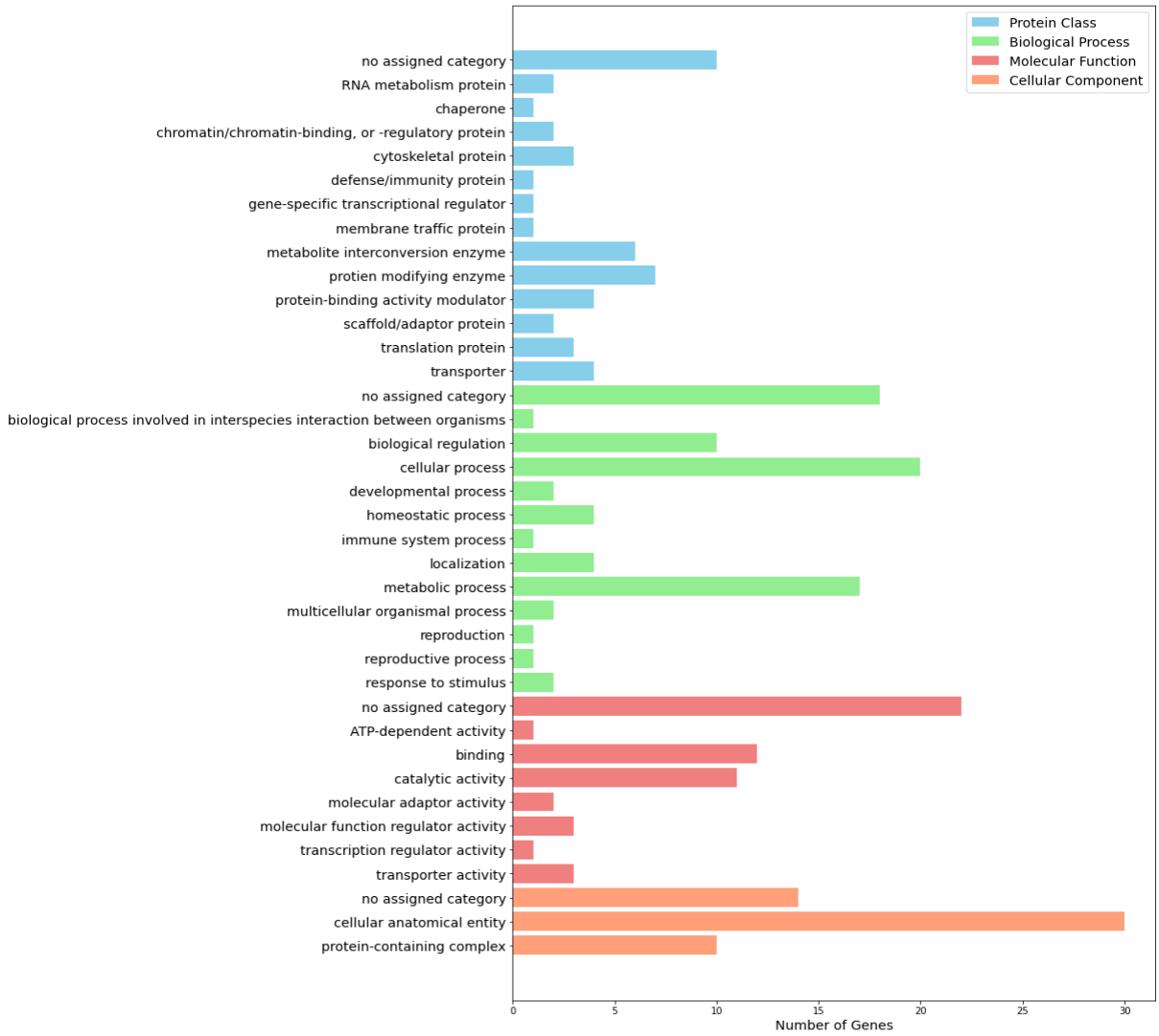

Fig. 15: Gene ontology (GO) enrichment analysis results based on the core set of coleoid-specific single-copy orthologous genes.

3.4. Assessing mollusk genome completeness

Since cephalopods belong to the phylum Mollusca, other molluscan genome assemblies can be used as a comparative framework to evaluate the performance of the cephalopod- and coleoid-specific BUSCO lineages. To this end, genomic data from three different bivalve species (*C. gigas*, *M. yessoensis*, *M. coruscus* and *A. kagoshimensis*) and two gastropod species (*C. squamiferum* and *A. immaculata*) were analyzed. Each species' genome was evaluated using the cephalopod-specific BUSCO lineage version 2 (n=512) and the coleoid-specific BUSCO lineage (n=746) to assess their conservation of orthologous genes.

The genome completeness of *A. immaculata* using the cephalopod-specific BUSCO lineage was C:46.9% [S:42.4%, D:4.5%], F:0.0%, M:53.1%. When using the coleoid-specific BUSCO lineage, the completeness dropped to C:27.5% [S:23.9%, D:3.6%], F:0.0%, M:72.5%. For the second gastropod species, *C. squamiferum*, the cephalopod-specific dataset yielded C:57.2% [S:56.6%, D:0.6%], F:0.0%, M:42.8%, while the coleoid-specific dataset returned C:27.5% [S:23.9%, D:3.6%], F:0.0%, M:72.5% (see **Fig. 16**).

In the case of the bivalve species, *A. kagoshimensis*, genome completeness was C:56.3% [S:50.4%, D:5.9%], F:0.0%, M:43.7% with the cephalopod-specific BUSCO lineage and C:31.3% [S:22.5%, D:8.8%], F:0.0%, M:68.7% with the coleoid-specific lineage. *M. coruscus* showed C:53.7% [S:53.3%, D:0.4%], F:0.0%, M:46.3% with the cephalopod-specific dataset and C:30.8% [S:30.0%, D:0.8%], F:0.0%, M:69.2% with the coleoid-specific dataset. The genome completeness for *M. yessoensis* was C:55.5% [S:55.1%, D:0.4%], F:0.0%, M:44.5% with the cephalopod-specific lineage and C:31.4% [S:31.1%, D:0.3%], F:0.0%, M:68.6% with the second dataset. Lastly, *C. gigas* showed C:53.0% [S:52.0%, D:1.0%], F:0.0%, M:47.0% with the cephalopod-specific BUSCO lineage and C:30.3% [S:29.1%, D:1.2%], F:0.0%, M:69.7% with the coleoid-specific BUSCO lineage (see **Fig. 16**).

As displayed in **Fig. 16**, the total numbers of single-copy (S), duplicated (D) and missing (M) genes are shown for each species. Since no fragmented orthologous genes were found in any of the genome assemblies, these were not included in the figure. The results, as well as **Fig. 16**, indicated that the cephalopod-specific dataset consistently identified fewer missing or duplicated orthologs and more single-copy orthologs across all species, despite containing less orthologous genes (n=512) compared to the coleoid-specific dataset (n=746). An exception was observed in *M. yessoensis*, where both datasets identified two duplicated orthologs, and in *C. squamiferum*, where the coleoid-specific lineage identified one less duplicated gene compared to the cephalopod-specific lineage (see. **Fig. 16**).

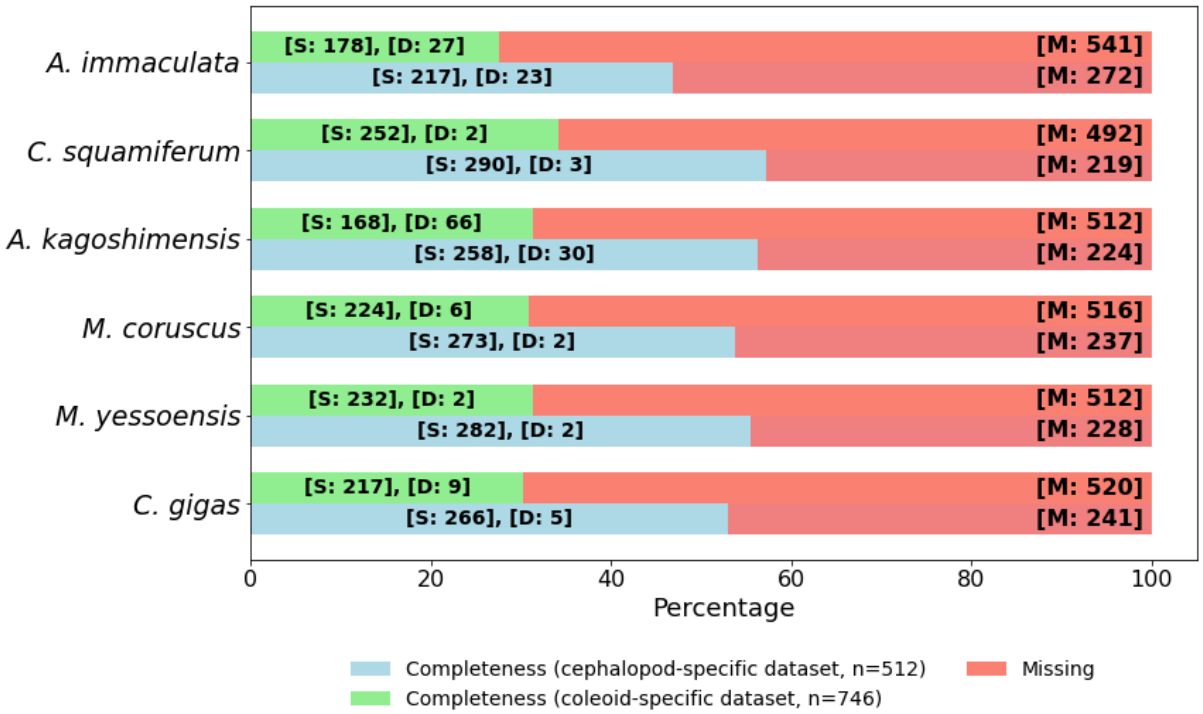

Fig. 16: Results of the cephalopod- and coleoid-specific BUSCO lineage analyses for various mollusk species. The bars indicate the percentage of complete and missing orthologs identified in each species. The numbers inside the brackets represent the total counts of single-copy (S), duplicated (D) and missing (M) genes.

4. Discussion

Making chromosome-scale (high-quality) genome assemblies available is crucial for several reasons. For instance, the assembly level significantly impacts the annotation of genomes and further downstream analyses, such as functional annotations and phylogenetic studies. High-quality assemblies are also essential for accurately assessing the number of repetitive elements in the genome. This is important because repetitive elements can interfere with gene predictions, potentially resulting in overestimating gene numbers through wrongful prediction of such elements as genes due to sequence similarities (Hoff & Stanke 2019). Therefore, high-quality assemblies are necessary to obtain accurate gene sets, as these sets are foundational for reliable genome annotations and subsequent research into genome structure, function and evolution.

4.1. Comparison of the cephalopod-specific sets of orthologs

The main difference between the cephalopod-specific BUSCO lineage version1 and version 2 lies in the source of the protein sequences. For version 1, we used previously published protein sequences of *E. scolopes*, *D. pealeii*, *O. bimaculoides* and *N. pompilius*, which were derived from gene modeling. In contrast, for version 2, we utilized the MIKADO v2.3.4 pipeline, which combines transcripts from various sources and methods (Venturini *et al.* 2018). For subsequent steps we used version 2 for several reasons that we will discuss in the following.

Programs for gene prediction, which are often used to derive sequences, rely on mathematical approaches of biological signals, like splice sites. There are two primary methods for gene prediction: *ab initio* and *similarity based*. The *ab initio* method involves training on a known set of genes to model biological signals and coding/non-coding regions. However, this method has limitations particularly with longer genomes, where its accuracy tends to decrease. On the other hand, the *similarity based* method uses external data such as homologous protein information but is also limited if homologous data is insufficient (Hoff & Stanke 2019; Stanke & Waack 2003).

Gene prediction tools, such as GeneMark-EP (Lomsadze 2005; Ter-Hovhannisyan *et al.* 2008) or Augustus (Hoff & Stanke 2019; Stanke & Waack 2003), generate so called “genomic seeds” – short sequences that serve as references for gene prediction algorithms. These seeds are typically conserved sequences or motifs resembling parts of known genes. These short seeds allow gene prediction tools to process large numbers of protein sequences more efficiently. However, when dealing with larger genomes, gene prediction accuracy can suffer, particularly during genomic seed generation (Brůna *et al.* 2023). Larger genomes often contain high

numbers of repeats, further complicating gene predictions. This complexity can lead to overestimating gene numbers and the prediction of shorter, fragmented, or incomplete genes (Brůna *et al*. 2023; Hoff & Stanke 2019).

Given these challenges with traditional gene prediction methods, the MIKADO v2.3.4 pipeline, used for version 2 provides a more reliable approach to gene annotation. Unlike conventional gene prediction programs that depend heavily on potentially inaccurate predictions, the MIKADO v2.3.4 pipeline combines transcripts from various sources and methods, providing a more robust basis for gene annotations. Throughout its multi-step process, the pipeline standardizes transcripts, removes duplications and artifacts, and consolidates them into a single database, grouping them into loci. During the combination step, it utilizes ORFs, protein similarity, and high-quality splice junctions to manage transcript fusions effectively. This approach has previously been successfully applied, for example, to the large and repeat-rich genome of *Triticum awstivum*, where it reduced the initial dataset by 30-fold, significantly improving the annotation quality and saving time in further downstream analyses (Venturini *et al*. 2018).

In the context of cephalopod genome analyses, using cephalopod RNA-seq data derived from Illumina or PacBio sequencing – both compatible with the MIKADO v2.3.4 pipeline – provides robust external evidence of gene expression, especially in specific organs. Focusing on transcript evidence allows for the reconstruction of long, complete ORFs, which are essential for accurately estimating genome completeness and fragmentation. Since cephalopods typically possess around 20,000 to 30,000 genes, relying solely on shorter predicted genes could lead to the unintended exclusion of many genes, resulting in less reliable outcomes. By using RNA-seq data and the MIKADO v2.3.4. pipeline, the potential pitfalls of short, predicted genes are minimized, leading to a more accurate evaluation of transcript completeness and integrity.

Although gene prediction tools are often assumed to produce shorter genes compared to annotation programs based on external evidence, this assumption was not confirmed in this study. The cephalopod-specific BUSCO lineage version 1, which is based on predicted genes, provided a larger mean orthogroup length and a higher number of predicted genes compared to version 2. However, these predicted gene numbers can be inflated due to the probabilistic nature of gene prediction models, which may overestimate gene counts based on sequence patterns. In contrast, version 2, which uses the MIKADO v2.3.4 pipeline, offered fewer but more reliable and accurate transcripts because it relied on real evidence rather than just probabilistic predictions.

Additionally, a higher percentage of orthologs were correctly mapped to chromosomes using the MIKADO v2.3.4 pipeline compared to gene prediction tools. Given the results, even though predicted genes from version 1 were longer and more numerous, the MIKADO v2.3.4. pipeline still produced more accurate and evidence-based transcripts, fulfilling its intended purpose as initially described by its authors, leading to the decision of progressing with the customized cephalopod-specific BUSCO lineage version 2 for further analyses.

4.2. Cephalopod-specific *versus* coleoid-specific BUSCO lineages

Since both the cephalopod-specific (version 2) and coleoid-specific BUSCO lineage were generated using the MIKADO v2.3.4 pipeline, there are no differences between predicted genes and evidence-based transcript annotations, as both relied on external evidence.

The two cephalopod subclasses, nautiloids and coleoids, diverged during the Cambrian era (541 – 485.4 mya), followed by a further split within the coleoids into Decapodiformes and Octopodiformes in the Permian era (270 mya) (Kröger *et al*. 2011; Sanchez *et al*. 2018; Tanner *et al*. 2017). Given that the division between nautiloids and coleoids occurred earlier than the split within coleoids, it is not surprising that a dataset based solely on closer related coleoid species identified higher numbers of shared single-copy orthologs than the one including nautiloid species. This difference can be attributed to the rapid evolution within coleoid cephalopods, leading to the emergence of novel organs and genes. Examples of such novel organs include the light organ in bobtail squids like *E. scolopes*, the accessory nidamental glands in squids, and the suckers in octopods (Ritschard *et al*. 2019).

In species that lack these organs, such as the ancient nautiloid *N. pompilius*, the corresponding novel genes are not expressed, resulting in no support from RNA-seq data (Brůna *et al*. 2023). Novelty, in this context, refers to the absence of similar sequences between species or taxa, and novel genes are those without homologous counterparts in other species (Klasberg *et al*. 2016; Ritschard *et al*. 2019). These genes can be derived from non-coding sequences or through rapid evolution following duplication events (reviewed in Ritschard *et al*. 2019). The presence of novel genes may explain why the coleoid-specific BUSCO lineage contained more single-copy orthologs, as it was generated using two squid and octopod species. In contrast, the cephalopod-specific gene set included *N. pompilius*, which lacks these organs and therefore these novel genes of coleoids, leading to fewer single-copy orthologs.

However, despite the lower number of single-copy orthologs, the gene set that includes *N. pompilius* is still preferred over the one based on solely coleoid species, because extant

coleoids can still possess internal calcified structures, such as the internal shell of the Ram's horn squid, the phragmocone of cuttlefish, or the reduced shell, known as the gladius, in squids (Sanchez *et al*. 2018). In these cases, *N. pompilius* therefore represents a good outgroup.

4.3. Comparison between MetaEuk and Miniprot

Over the course of this study, several updates for BUSCO were released, specifically with the release of version v5.5.0 and beyond (Simão *et al*. 2015). One notable advancement is the integration of the protein-to-genome alignment program Miniprot (Li 2023), which became available with BUSCO v5.5.0. Miniprot was designed to enhance the speed and accuracy of protein-to-genome alignments, surpassing previous aligners in both performance and efficiency (Li 2023). From BUSCO v5.5.0 onward, Miniprot had been adopted as the default aligner for the eukaryotic database (eukaryota_odb10), though other BUSCO lineages still offer the option to switch from the initial aligner, MetaEuk (Levy Karin *et al*. 2020).

The results of this study corroborate Miniprot's claimed advantage. Specifically, it demonstrated higher accuracy compared to MetaEuk, as evidenced by the improved annotation of genes to chromosomes within the cephalopod-specific dataset. The time-saving benefits of Miniprot further justified its primary use in conjunction with BUSCO v5.5.0 for this study.

Additionally, the authors of Miniprot have developed a similar tool called compleasm, which they argue offers even greater efficiency and accuracy than BUSCO. Compleasm requires only a single round of alignment with Miniprot instead of BUSCO's two rounds of MetaEuk (initial- and re-run) (Huang & Li 2023). However, during this study, compleasm could not be utilized with the cephalopod-specific BUSCO lineage version 2, due to its restriction to pre-defined BUSCO lineages. Consequently, testing compleasm was not feasible within this study. Future updates might allow for custom lineages with compleasm, potentially integrating the cephalopod-specific dataset and further validating its performance.

4.4. BUSCO lineage comparisons with other datasets

Compared to the cephalopod-specific BUSCO lineage version 2, other BUSCO datasets did not perform well with cephalopod genomes.

Given the fact that cephalopod genomes are very large, highly repetitive, and have significant intron sizes (Albertin *et al*. 2012; Kim *et al*. 2018; Schmidbaur *et al*. 2022), as well as the emergence of novel organs within coleoids (Ritschard *et al*. 2019), it was expected that the

eukaryotic (eukaryota_odb10) and metazoan (metazoa_odb10) BUSCO lineages would not perform well with such complex genomes. Surprisingly, the molluscan (mollusca_odb10) dataset performed the worst in resolving the completeness of cephalopod genome assemblies. This outcome may be largely attributed to the presence of novel organs in cephalopods or specifically in coleoids, which contain novel genes not shared with other eukaryotic or metazoan species. As a result, genes associated with these tissues and organs might lack RNA-seq evidence due to their absence in the broader molluscan lineage. Additionally, cephalopods have undergone large gene family expansions, further contributing to the complexity of their genomes (Albertin *et al*. 2015; reviewed in Albertin & Simakov 2020), which may be another reason for the poor performance of the other datasets.

Besides the presence of novel genes in cephalopods, the composition of the eukaryotic, metazoan, and molluscan datasets reveals that they are not adequately suited for assessing cephalopod genomes. The eukaryotic dataset includes 70 different species, of which 38 belong to Metazoa, with only a single mollusk, the bivalve *Lottia gigantea*, represented. The metazoan BUSCO dataset includes 65 species, with only two mollusks – *C. gigas* and *L. gigantea* – both bivalves. Most species in this dataset belong to arthropods (33 representatives), followed by 20 vertebrate species. The molluscan dataset includes seven different mollusk species: three bivalves (*C. gigas*, *L. gigantea* and *M. yessoensis*), two gastropods (*Biomphalaria glabrata* and *Aplysia californica*), and one cephalopod (*O. bimaculoides*) (see List of BUSCO.v4 lineages (ezlab.org).

Including one more mollusk in the metazoan dataset might explain why it performed better in terms of cephalopod genome completeness compared to the eukaryotic dataset. Additionally, the higher representation of vertebrates in the metazoan dataset, as opposed to the eukaryotic dataset with only elven vertebrate species, could contribute to its better performance. Vertebrates have large, complex genomes, often resulting from whole-genome duplications, a characteristic they share with cephalopods (Albertin *et al*. 2015; Ritschard *et al*. 2019). Some conserved genes and gene structures, such as Hox genes, are assumed to have been present in the last common ancestor of all metazoans (reviewed in (Ritschard *et al*. 2019). Although whole-genome duplications in vertebrates and cephalopods would have occurred independently, they might share more conserved genes. However, to avoid speculation, further in-depth analyses of the shared genes between these groups are necessary, especially because the large size of cephalopod genomes are attributed more to whole-genome rearrangements rather than duplications (Albertin *et al*. 2015; reviewed in Albertin & Simakov 2020; Belcaid *et al*. 2019; Ritschard *et al*. 2019).

Although it was anticipated that the eukaryotic and metazoan datasets might not perform optimally with cephalopod genomes, it was unexpected that the molluscan dataset produced the worst results among all BUSCO lineages used in this study. Even though the Venn diagrams showed that the cephalopod-specific and coleoid-specific datasets shared most of their genes with the molluscan dataset, the poor performance of the molluscan dataset underscores its unsuitability for cephalopods. This is also seen in other studies, such as in Destanović *et al*. (2023), where they improved the genome of *O. vulgaris* to chromosome-level and achieved better completeness with the metazoan BUSCO lineage compared to the molluscan lineage. This inadequacy likely stems from the fact that only one cephalopod species was included in generating the molluscan-specific set of orthologs. This limitation is further exacerbated by the evolution of novel organs within cephalopods, which are not represented in other mollusk groups.

The results of this study, therefore, emphasize the necessity of using a suitable dataset for the complex, repeat-rich, and large genome of cephalopods. Although the metazoan dataset can provide a preliminary estimate of genome completeness, a more tailored approach is required for accurate and reliable assessments.

4.5. Gene orthology

Genes are organized on chromosomes, and their specific locations can vary between species and taxa (Dixon *et al*. 2012). Cephalopods exhibit significant variation in their karyotypes, with nautiloids typically exhibiting a diploid chromosome set of 52 (n=26), Octopodiformes of 60 (n=30), and Decapodiformes of 92 (n=46) (Wang & Zheng 2017). The reorganization of genes on chromosomes can occur through chromosomal fission or fusion, which likely explains the large differences in chromosome numbers within cephalopods (Albertin *et al*. 2015, 2022).

In this study, comparisons between Decapodiformes species confirmed the commonly assumed number of 46 chromosomes (2n=92) in squids, though exceptions were observed. For example, potential chromosomal fusion events were identified when comparing *E. scolopes* and *D. pealeii*, where chromosomes 24 and 40 in *D. pealeii* may correspond to chromosome 2 in *E. scolopes*. This potential fusion suggests the relocation of cephalopod-specific single-copy orthologs in *E. scolopes*. Interestingly, such a chromosomal fusion was not detected using the coleoid-specific gene set, potentially due to the substitution of *N. pompilius* with *O. vulgaris* in the dataset. While there might be no relocation of coleoid-specific genes, cephalopod-specific genes could lead to chromosomal rearrangements. However, both gene sets showed an ortholog distribution across all chromosomes, highlighting the

importance of including *N. pompilius* as an outgroup for coleoids in selecting an appropriate cephalopod-specific single-copy orthologous gene set.

Potential chromosomal fusions were detected in *S. atlantica* with both single-copy orthologous gene sets. Additionally, only 41 chromosomes were assembled in *S. atlantica*, which might indicate fusion events or possible misassemblies. These observations raise whether orthologs involved in translocations or fusions evolve more rapidly, warranting further research.

To improve comparisons between squid species, it might be advantageous to standardize chromosome naming based on well-assembled genomes, such as those of *E. scolopes* or *D. pealeii*. A useful approach could be to rank chromosomes by sequence length, similar to human chromosomes, where the first chromosome is the largest, with a length three times greater than chromosome 22 (National Center for Biotechnology Information (US) 1998).

4.6. Gene ontology

Cephalopods have undergone extensive gene family expansions (reviewed in Albertin & Simakov 2020), and tools like Gene Ontology (GO) enrichment analysis are valuable for identifying and categorizing different gene sets. However, the GO analysis conducted on the cephalopod- and coleoid-specific ortholog sets in this study did not reveal any significant differences between them. Notably, a large portion of the genes could not be assigned to a protein class, biological process, or molecular function. This suggests that many of these uncharacterized genes may be unique to cephalopods, or more specifically to coleoids. To confirm this hypothesis and better understand these genes, further research is required.

4.7. Mollusk genome completeness assessment

Three different bivalve and two different gastropod species showed varying results when their genome completeness was estimated using either the cephalopod-specific BUSCO lineage version 2 or the coleoid-specific BUSCO lineage. The cephalopod-specific dataset included *N. pompilius*, an ancient cephalopod species with an external calcified shell. Given that coleoid cephalopods typically have a reduced or absent shell, it was anticipated that shelled mollusks would exhibit higher genome completeness when assessed with the cephalopod-specific dataset.

As discussed in chapter 4.2. coleoids have undergone rapid evolution, leading to the development of novel organs not found in nautiloids or other mollusks (Ritschard *et al.* 2019).

This rapid evolution likely explains the higher number of missing genes when using the coleoid-specific gene set for genome completeness assessment.

Furthermore, because *N. pompilius* retains an external shell, it expresses genes involved in biomineralization. Previous studies have identified a conserved "toolkit" responsible for biomineralization across shelled mollusks (Kocot *et al.* 2016; Zhang *et al.* 2021). These genes are typically expressed in the mantle, which secretes the shell, but this tissue is also rapidly evolving, resulting in variations in shell structures, colors, or patterns among closely related species and taxa (Kocot *et al.* 2016).

Since biomineralization genes evolve rapidly and have undergone duplication events (Kocot *et al.* 2016), it is possible that these genes are present in coleoids, which also possess a mantle – a key morphological feature shared by all mollusks. Given the divergence between nautiloids and coleoids around 270 mya, and the rapid evolution within coleoids (Kröger *et al.* 2011; Sanchez *et al.* 2018; Tanner *et al.* 2017), the biomineralization genes in these species may no longer perform their original functions, potentially leading to their exclusion from the conserved "toolkit" present in shelled mollusks.

Despite this, *N. pompilius* still shares the conserved biomineralization "toolkit" with other mollusks (Zhang *et al.* 2021), which likely contributes to the higher genome completeness observed in the cephalopod-specific gene set when assessing bivalve or gastropod genomes. Furthermore, other studies have shown that *N. pompilius* exhibits greater macrosynteny conservation, which describes the gene order conservation at the chromosomal level (Albertin & Simakov 2020), with other mollusks than with coleoids (Albertin *et al.* 2022). This conservation between nautiloids and other mollusks may also explain the higher completeness observed with the cephalopod-specific dataset. This reinforces the importance of considering evolutionary context when selecting gene sets for genome completeness assessments.

4.8. Conclusion

Overall, the cephalopod-specific BUSCO lineage version 2, processed with the MIKADO v2.3.4 pipeline, emerged as the most suitable for assessing cephalopod genomes among all customized BUSCO lineages investigated in the present study. This finding is particularly significant because this dataset provides a more accurate set of genes for cephalopod genome assessments compared to other BUSCO lineages, including the eukaryotic, metazoan or molluscan lineages. Such accuracy underscores the necessity of a cephalopod-specific set of single-copy orthologs for genome completeness assessments, which can be utilized in tools

like BUSCO or, in the future, compleasm. Beyond completeness assessments, this gene set also holds promise for other research areas, such as studying chromosomal fusion and fission events.

Furthermore, the inclusion of *N. pompilius* is crucial due to its evolutionary distance from coleoids, which enhances the accuracy of completeness assessments. Understanding this relationship is vital for understanding gene family expansions in cephalopods and coleoids, in contrast to other mollusks and metazoans.

Finally, the pipeline implemented in this study, MIKADO v2.3.4, can significantly improve future genome assemblies, not only in cephalopod species but across a broad range of taxa.

5. Future perspectives

In this project, we generated two core sets of single-copy orthologous genes: a cephalopod-specific set derived from three coleoid species (*E. scolopes*, *D. pealeii*, and *O. bimaculoides*) and one nautiloid species (*N. pompilius*), as well as a coleoid-specific set where *N. pompilius* was substituted with *O. vulgaris*. These gene sets, designed for use with BUSCO, provide a foundation for more accurate genome completeness assessments, and their utility extends beyond the cephalopods to other metazoans. The MIKADO pipeline, which leverages RNA-seq data, has demonstrated its superiority over gene-model based annotation data in this project. This approach not only improved annotations for cephalopods but also holds promise for enhancing genome assemblies across a broad range of taxa.

Looking ahead, future research can build on these core gene sets by delving deeper into the synteny of cephalopods, exploring their roles in chromosomal rearrangements, and investigating the evolution of uncharacterized genes. These studies could shed light on the unique genomic architecture of cephalopods, contributing to a deeper understanding of their biology and evolution.

Additionally, the comparative analysis of synteny between different cephalopod species using these gene sets has the potential to reveal new insights into the mechanisms of gene family expansions and chromosomal dynamics in these organisms. Expanding this research to include a broader range of cephalopod species and other mollusks could help refine the BUSCO lineage and further improve its accuracy.

Furthermore, as the Earth Biogenome Project (EBP) progresses towards its goal of sequencing and annotating all eukaryotic life by 2028 (Lewin *et al.* 2018), the methodologies and findings from this study could play a role in achieving high-quality genome assemblies for a diverse

array of species. Continued refinement of these tools and approaches will be essential for meeting the challenges posed by the complexity of cephalopod genomes and other understudied taxa.

In conclusion, our work provides a useful foundation for future projects, but much remains to be done. Future research should focus on expanding the scope of these gene sets, exploring their functional implications, and applying these findings to a wider array of species. This will not only advance our understanding of cephalopod genomics but also contribute to the broader goals of genomic research and biodiversity conservation.

References

Albertin, C.B., Bonnaud, L., Brown, C.T., Crookes-Goodson, W.J., Da Fonseca, R.R., Di Cristo, C., Dilkes, B.P., Edsinger-Gonzales, E., Freeman, R.M., Hanlon, R.T., Koenig, K.M., Lindgren, A.R., Martindale, M.Q., Minx, P., Moroz, L.L., Nödl, M.-T., Nyholm, S.V., Ogura, A., Pungor, J.R., Rosenthal, J.J.C., Schwarz, E.M., Shigeno, S., Strugnell, J.M., Wollesen, T., Zhang, G. & Ragsdale, C.W. (2012) Cephalopod genomics: A plan of strategies and organization. *Standards in Genomic Sciences* 7, 175–188. https://doi.org/10.4056/sigs.3136559

Albertin, C.B., Medina-Ruiz, S., Mitros, T., Schmidbaur, H., Sanchez, G., Wang, Z.Y., Grimwood, J., Rosenthal, J.J.C., Ragsdale, C.W., Simakov, O. & Rokhsar, D.S. (2022) Genome and transcriptome mechanisms driving cephalopod evolution. *Nature Communications* 13, 2427. https://doi.org/10.1038/s41467-022-29748-w

Albertin, C.B. & Simakov, O. (2020) Cephalopod Biology: At the Intersection Between Genomic and Organismal Novelties. *Annual Review of Animal Biosciences* 8, 71–90. https://doi.org/10.1146/annurev-animal-021419-083609

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W. & Rokhsar, D.S. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* 524, 220–224. https://doi.org/10.1038/nature14668

Allcock, A.L., Lindgren, A. & Strugnell, J.M. (2015) The contribution of molecular data to our understanding of cephalopod evolution and systematics: a review. *Journal of Natural History* 49, 1373–1421. https://doi.org/10.1080/00222933.2013.825342

Belcaid, M., Casaburi, G., McAnulty, S.J., Schmidbaur, H., Suria, A.M., Moriano-Gutierrez, S., Pankey, M.S., Oakley, T.H., Kremer, N., Koch, E.J., Collins, A.J., Nguyen, H., Lek, S., Goncharenko-Foster, I., Minx, P., Sodergren, E., Weinstock, G., Rokhsar, D.S., McFall-Ngai, M., Simakov, O., Foster, J.S. & Nyholm, S.V. (2019) Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *Proceedings of the National Academy of Sciences* 116, 3030–3035. https://doi.org/10.1073/pnas.1817322116

Berthold, T. & Engeser, T. (1987) Phylogenetic analysis and systematization of the Cephalopoda (Mollusca). *Verhandlungen des Naturwissenschaftlichen Vereins in Hamburg (Neue Folge)* 29, 187–220.

Boletzky, S. v. (2003) Biology of Early Life Stages in Cephalopod Molluscs. In: *Advances in Marine Biology*. Elsevier, pp. 143–203.

Bonnaud, L., Ozouf-Costaz, C. & Boucher-Rodoni, R. (2004) A molecular and karyological approach to the taxonomy of Nautilus. *Comptes Rendus. Biologies* 327, 133–138. https://doi.org/10.1016/j.crvi.2003.12.004

Bradnam, K. (2015) Goodbye CEGMA, hello BUSCO! Available from: Goodbye CEGMA, hello BUSCO! — ACGT

Brůna, T., Li, H., Guhlin, J., Honsel, D., Herbold, S., Stanke, M., Nenasheva, N., Ebel, M., Gabriel, L. & Hoff, K.J. (2023) Galba: genome annotation with miniprot and AUGUSTUS. *BMC Bioinformatics* 24, 327. https://doi.org/10.1186/s12859-023-05449-z

Buchfink, B., Xie, C. & Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12, 59–60. https://doi.org/10.1038/nmeth.3176

Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Colgan, J.A. (2020) The Power of The Cowrie Shell: Human Currency. https://scholar.colorado.edu/concern/graduate_thesis_or_dissertations/t722h999w.

da Fonseca, R.R., Couto, A., Machado, A.M., Brejova, B., Albertin, C.B., Silva, F., Gardner, P., Baril, T., Hayward, A., Campos, A., Ribeiro, Â.M., Barrio-Hernandez, I., Hoving, H.-J., Tafur-Jimenez, R., Chu, C., Frazão, B., Petersen, B., Peñaloza, F., Musacchia, F., Alexander, G.C., Osório, H., Winkelmann, I., Simakov, O., Rasmussen, S., Rahman, M.Z., Pisani, D., Vinther, J., Jarvis, E., Zhang, G., Strugnell, J.M., Castro, L.F.C., Fedrigo, O., Patricio, M., Li, Q., Rocha, S., Antunes, A., Wu, Y., Ma, B., Sanges, R., Vinar, T., Blagoev, B., Sicheritz-Ponten, T., Nielsen, R. & Gilbert, M.T.P. (2020) A draft genome sequence of the elusive giant squid, Architeuthis dux. *GigaScience* 9, giz152. https://doi.org/10.1093/gigascience/giz152

Destanović, D., Schultz, D.T., Styfhals, R., Cruz, F., Gómez-Garrido, J., Gut, M., Gut, I., Fiorito, G., Simakov, O., Alioto, T.S., Ponte, G. & Seuntjens, E. (2023) A chromosome-level reference genome for the common octopus, *Octopus vulgaris* (Cuvier, 1797) P. Campbell (Ed). *G3: Genes, Genomes, Genetics* 13, jkad220. https://doi.org/10.1093/g3journal/jkad220

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. & Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380. https://doi.org/10.1038/nature11082

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dolinski, K. & Botstein, D. (2007) Orthology and Functional Conservation in Eukaryotes. *Annual Review of Genetics* 41, 465–507. https://doi.org/10.1146/annurev.genet.40.110405.090439

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B.L., Soler, L., Binzer-Panchal, M. & Lantz, H. (2018) Ten steps to get started in Genome Assembly and Annotation. *F1000Research* 7, 148. https://doi.org/10.12688/f1000research.13598.1

Eddy, S.R. (2011) Accelerated Profile HMM Searches W. R. Pearson (Ed). *PLoS Computational Biology* 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195

El Hilali, S. & Copley, R.R. (2023) macrosyntR : Drawing automatically ordered Oxford Grids from standard genomic files in R.

Emms, D.M. & Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, 157. https://doi.org/10.1186/s13059-015-0721-2

Farrington, J.W., Tripp, B.W., Tanabe, S., Subramanian, A., Sericano, J.L., Wade, T.L. & Knap, A.H. (2016) Edward D. Goldberg's proposal of "the Mussel Watch": Reflections after 40 years. *Marine Pollution Bulletin* 110, 501–510. https://doi.org/10.1016/j.marpolbul.2016.05.074

Figueras, A., Moreira, R., Sendra, M. & Novoa, B. (2019) Genomics and immunity of the Mediterranean mussel Mytilus galloprovincialis in a changing environment. *Fish & Shellfish Immunology* 90, 440–445. https://doi.org/10.1016/j.fsi.2019.04.064

Gabaldón, T. & Koonin, E.V. (2013) Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* 14, 360–366. https://doi.org/10.1038/nrg3456

Gavriouchkina, D., Tan, Y., Ziadi-Künzli, F., Hasegawa, Y., Piovani, L., Zhang, L., Sugimoto, C., Luscombe, N., Marlétaz, F. & Rokhsar, D.S. (2022) A single-cell atlas of bobtail squid visual and nervous system highlights molecular principles of convergent evolution.

Gazeau, F., Parker, L.M., Comeau, S., Gattuso, J.-P., O'Connor, W.A., Martin, S., Pörtner, H.-O. & Ross, P.M. (2013) Impacts of ocean acidification on marine shelled molluscs. *Marine Biology* 160, 2207–2245. https://doi.org/10.1007/s00227-013-2219-3

Gomes-dos-Santos, A., Lopes-Lima, M., Castro, L.F.C. & Froufe, E. (2020) Molluscan genomics: the road so far and the way forward. *Hydrobiologia* 847, 1705–1726. https://doi.org/10.1007/s10750-019-04111-1

Haas, B.J. (2023) TransDecoder Release v5.7.1.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N. & Regev, A. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8, 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hallinan, N.M. & Lindberg, D.R. (2011) Comparative Analysis of Chromosome Counts Infers Three Paleopolyploidies in the Mollusca. *Genome Biology and Evolution* 3, 1150–1163. https://doi.org/10.1093/gbe/evr087

Hanbo Chen (2011) VennDiagram: Generate High-Resolution Venn and Euler Plots. , 1.7.3.

Hanlon, R.T. & McManus, G. (2020) Flamboyant cuttlefish behavior: Camouflage tactics and complex colorful reproductive behavior assessed during field studies at Lembeh Strait, Indonesia. *Journal of Experimental Marine Biology and Ecology* 529, 151397. https://doi.org/10.1016/j.jembe.2020.151397

Hanlon, R.T. & Messenger, J.B. (2018) *Cephalopod Behaviour*. 2nd ed. Cambridge University Press. Available from: https://www.cambridge.org/core/product/identifier/9780511843600/type/book (May 15, 2024)

Harrington, B., MenTaLguY, Hurst, N. & Gould, T. (2003) Inkscape.

Haszprunar, G. & Wanninger, A. (2012) Molluscs. *Current Biology* 22, R510–R514. https://doi.org/10.1016/j.cub.2012.05.039

Hoff, K.J. & Stanke, M. (2019) Predicting Genes in Single Genomes with AUGUSTUS. *Current Protocols in Bioinformatics* 65, e57. https://doi.org/10.1002/cpbi.57

Huang, N. & Li, H. (2023) compleasm: a faster and more accurate reimplementation of BUSCO T. Marschall (Ed). *Bioinformatics* 39, btad595. https://doi.org/10.1093/bioinformatics/btad595

Hunter, J.D. (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Jereb, P. & Roper, C.F.E. eds. (2005) *Cephalopods of the world: an annotated and illustrated catalogue of cephalopod species known to date*. Food and Agriculture Organization of the United Nations, Rome, 1 p.

Jereb, P., Roper, C.F.E., Norman, M.D., Finn, J.K., & FAO eds. (2005) *Octopods and vampire squids*. Entirely rewritten, revised and updated version. Food and Agriculture Organization of the United Nations, Rome, 352 pp.

Katoh, K. & Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kim, B.-M., Kang, S., Ahn, D.-H., Jung, S.-H., Rhee, H., Yoo, J.S., Lee, J.-E., Lee, S., Han, Y.-H., Ryu, K.-B., Cho, S.-J., Park, H. & An, H.S. (2018) The genome of common long-arm octopus Octopus minor. *GigaScience*. https://doi.org/10.1093/gigascience/giy119

Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Klasberg, S., Bitard-Feildel, T. & Mallet, L. (2016) Computational Identification of Novel Genes: Current and Future Perspectives. *Bioinformatics and Biology Insights* 10, BBI.S39950. https://doi.org/10.4137/BBI.S39950

Kocot, K.M., Aguilera, F., McDougall, C., Jackson, D.J. & Degnan, B.M. (2016) Sea shell diversity and rapidly evolving secretomes: insights into the evolution of biomineralization. *Frontiers in Zoology* 13, 23. https://doi.org/10.1186/s12983-016-0155-z

Koonin, E.V. (2005) Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* 39, 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Kröger, B., Vinther, J. & Fuchs, D. (2011) Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. *BioEssays* 33, 602–613. https://doi.org/10.1002/bies.201100001

Levy Karin, E., Mirdita, M. & Söding, J. (2020) MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8, 48. https://doi.org/10.1186/s40168-020-00808-x

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., Goldstein, M.M., Grigoriev, I.V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A., Richards, S., Castilla-Rubio, J.C., Van Sluys, M.-A., Soltis, P.S., Xu, X., Yang, H. & Zhang, G. (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* 115, 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, H. (2023) Protein-to-genome alignment with miniprot A. Valencia (Ed). *Bioinformatics* 39, btad014. https://doi.org/10.1093/bioinformatics/btad014

Lindgren, A.R. & Anderson, F.E. (2018) Assessing the utility of transcriptome data for inferring phylogenetic relationships among coleoid cephalopods. *Molecular Phylogenetics and Evolution* 118, 330–342. https://doi.org/10.1016/j.ympev.2017.10.004

Lomsadze, A. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* 33, 6494–6506. https://doi.org/10.1093/nar/gki937

Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. (2018) Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience* 7. https://doi.org/10.1093/gigascience/giy131

McFall-Ngai, M. (2008) Hawaiian bobtail squid. *Current Biology*, R1043–R1044.

National Center for Biotechnology Information (US). (1998) Genes and Disease [Internet] - Chromosome Map.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M. & Iyer, M.K. (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nature Methods* 14, 68–70. https://doi.org/10.1038/nmeth.4078

Packard, A. (1972) CEPHALOPODS AND FISH: THE LIMITS OF CONVERGENCE. *Biological Reviews* 47, 241–307. https://doi.org/10.1111/j.1469-185X.1972.tb00975.x

Parra, G., Bradnam, K. & Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. & Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33, 290–295. https://doi.org/10.1038/nbt.3122

Ponder, W.F. & Lindberg, D.R. eds. (2008) *Phylogeny and evolution of the Mollusca*. University of California Press, Berkeley, 469 pp.

Posit team (2024) RStudio: Integrated Development Environment for R.

Python Software Foundation (2001) Python Software Foundation.

R Core Team (2024) R: A Language and Environment for Statistical Computing.

Raybaut, P. & Cordoba, C. (2009) The Spyder Project Contributors.

Ritschard, E.A., Whitelaw, B., Albertin, C.B., Cooke, I.R., Strugnell, J.M. & Simakov, O. (2019) Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods: Co-evolutionary Signatures Across Levels of Genome Organization May Shed Light on Functional Linkage and Origin of Cephalopod Novelties. *BioEssays* 41, 1900073. https://doi.org/10.1002/bies.201900073

Sanchez, G., Jolly, J., Reid, A., Sugimoto, C., Azama, C., Marlétaz, F., Simakov, O. & Rokhsar, D.S. (2019) New bobtail squid (Sepiolidae: Sepiolinae) from the Ryukyu islands revealed by molecular and morphological analysis. *Communications Biology* 2, 465. https://doi.org/10.1038/s42003-019-0661-6

Sanchez, G., Setiamarga, D.H.E., Tuanapaya, S., Tongtherm, K., Winkelmann, I.E., Schmidbaur, H., Umino, T., Albertin, C., Allcock, L., Perales-Raya, C., Gleadall, I., Strugnell, J.M., Simakov, O. & Nabhitabhata, J. (2018) Genus-level phylogeny of cephalopods using

molecular markers: current status and problematic areas. *PeerJ* 6, e4331. https://doi.org/10.7717/peerj.4331

Schmidbaur, H., Kawaguchi, A., Clarence, T., Fu, X., Hoang, O.P., Zimmermann, B., Ritschard, E.A., Weissenbacher, A., Foster, J.S., Nyholm, S.V., Bates, P.A., Albertin, C.B., Tanaka, E. & Simakov, O. (2022) Emergence of novel cephalopod gene regulation and expression through large-scale genome reorganization. *Nature Communications* 13, 2172. https://doi.org/10.1038/s41467-022-29694-7

Sedlazeck, F.J., Lee, H., Darby, C.A. & Schatz, M.C. (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* 19, 329–346. https://doi.org/10.1038/s41576-018-0003-4

Shigeno, S., Parnaik, R., Albertin, C.B. & Ragsdale, C.W. (2015) Evidence for a cordal, not ganglionic, pattern of cephalopod brain neurogenesis. *Zoological Letters* 1, 26. https://doi.org/10.1186/s40851-015-0026-z

Sigwart, J.D., Lindberg, D.R., Chen, C. & Sun, J. (2021) Molluscan phylogenomics requires strategically selected genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376, 20200161. https://doi.org/10.1098/rstb.2020.0161

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sohn, J. & Nam, J.-W. (2016) The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*, bbw096. https://doi.org/10.1093/bib/bbw096

Sonnhammer, E.L.L. & Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics* 18, 619–620. https://doi.org/10.1016/S0168-9525(02)02793-2

Stanke, M. & Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Studer, R.A. & Robinson-Rechavi, M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics* 25, 210–216. https://doi.org/10.1016/j.tig.2009.03.004

Talmage, S.C. & Gobler, C.J. (2010) Effects of past, present, and future ocean carbon dioxide concentrations on the growth and survival of larval shellfish. *Proceedings of the National Academy of Sciences* 107, 17246–17251. https://doi.org/10.1073/pnas.0913804107

Tanner, A.R., Fuchs, D., Winkelmann, I.E., Gilbert, M.T.P., Pankey, M.S., Ribeiro, Â.M., Kocot, K.M., Halanych, K.M., Oakley, T.H., Da Fonseca, R.R., Pisani, D. & Vinther, J. (2017) Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution. *Proceedings of the Royal Society B: Biological Sciences* 284, 20162818. https://doi.org/10.1098/rspb.2016.2818

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. & Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research* 18, 1979–1990. https://doi.org/10.1101/gr.081612.108

The pandas development team (2024) pandas-dev/pandas: Pandas.

Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L. & Mi, H. (2022) PANTHER : Making genome-scale phylogenetics accessible to all. *Protein Science* 31, 8–22. https://doi.org/10.1002/pro.4218

Vandepas, L.E., Dooley, F.D., Barord, G.J., Swalla, B.J. & Ward, P.D. (2016) A revisited phylogeography of *Nautilus pompilius*. *Ecology and Evolution* 6, 4924–4935. https://doi.org/10.1002/ece3.2248

Venturini, L., Caim, S., Kaithakottil, G.G., Mapleson, D.L. & Swarbreck, D. (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* 7. https://doi.org/10.1093/gigascience/giy093

Wang, J. & Zheng, X. (2017) Comparison of the genetic relationship between nine Cephalopod species based on cluster analysis of karyotype evolutionary distance. *Comparative Cytogenetics* 11, 477–494. https://doi.org/10.3897/compcytogen.v11i3.12752

Wickham, H. (2016) *ggplot2: elegant graphics for data analysis*. Second edition. Springer, Switzerland, 260 pp.

Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. (2023) dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr.

Wickham, H., Hester, J., Chang, W. & Bryan, J. (2022) devtools: Tools to Make Developing R Packages Easier. https://devtools.r-lib.org/, https://github.com/r-lib/devtools.

Wu, T.D. & Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Young, R.E., Vecchione, M. & Donovan, D.T. (1998) The evolution of coleoid cephalopods and their present biodiversity and ecology. *South African Journal of Marine Science* 20, 393–420. https://doi.org/10.2989/025776198784126287

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P.W.H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., Peng, C., Meng, J., Yang, L., Liu, J., Wen, B., Zhang, N., Huang, Z., Zhu, Q., Feng, Y., Mount, A., Hedgecock, D., Xu, Z., Liu, Y., Domazet-Lošo, T., Du, Y., Sun, X., Zhang, S., Liu, B., Cheng, P., Jiang, X., Li, J., Fan, D., Wang, W., Fu, W., Wang, T., Wang, B., Zhang, J., Peng, Z., Li, Y., Li, N., Wang, J., Chen, M., He, Y., Tan, F., Song, X., Zheng, Q., Huang, R., Yang, H., Du, X., Chen, L., Yang, M., Gaffney, P.M., Wang, S., Luo, L., She, Z., Ming, Y., Huang, W., Zhang, S., Huang, B., Zhang, Y., Qu, T., Ni, P., Miao, G., Wang, J., Wang, Q., Steinberg, C.E.W., Wang, H., Li, N., Qian, L., Zhang, G., Li, Y., Yang, H., Liu, X., Wang, J., Yin, Y. & Wang, J. (2012) The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490, 49–54. https://doi.org/10.1038/nature11413

Zhang, Y., Mao, F., Mu, H., Huang, M., Bao, Y., Wang, L., Wong, N.-K., Xiao, S., Dai, H., Xiang, Z., Ma, M., Xiong, Y., Zhang, Z., Zhang, L., Song, X., Wang, F., Mu, X., Li, J., Ma, H., Zhang, Y., Zheng, H., Simakov, O. & Yu, Z. (2021) The genome of Nautilus pompilius illuminates eye evolution and biomineralization. *Nature Ecology & Evolution* 5, 927–938. https://doi.org/10.1038/s41559-021-01448-6

Zullo, L. & Hochner, B. (2011) A new perspective on the organization of an invertebrate brain. *Communicative & Integrative Biology* 4, 26–29. https://doi.org/10.4161/cib.13804