ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PhD programme in
Innovative technologies and sustainable use of Mediterranean Sea fishery and biological resources (FishMed-PhD)

Cycle XXXVII

Settore Consorsuale: 05/B1 – ZOOLOGIA E ANTROPOLOGIA

Settore Scientifico Disciplinare: BIO/05 – ZOOLOGIA

A comparative and evolutionary approach to study bivalve sex determination from a broad-phylogenetic perspective

Candidate: Filippo Nicolini

PhD Coordinator:
prof. Stefano Goffredo

Supervisor:
prof. Andrea Luchetti

Final exam 2025

Contents

List of abbreviations

AASD	amino acid sequence divergence
BSA	bovine serum albumin
CMS	cytoplasmatic male sterility
CUE	coupling of ubiquitin conjugation to endoplasmic reticulum degradation [domain]
DEAD/DEAH-box	Asp-Glu-Ala-Asp/Asp-Glu-Ala-His box
DGE	differential gene expression
DM	*dsx* and *mab-3* [domain]
DMA	DM-associated [domain]
Dmrt	*doublesex* and *mab-3* related transcription factor
Dmrt-1L	*doublesex and mab-3 related transcription factor 1-like*
Dm-W	*doublesex and mab-3 related gene W*
Dmy	*doublesex and mab-3 related gene Y*
DSFG	Dmrt, Sox, and Fox gene
dsx	*doublesex*
DUI	doubly uniparental inheritance
EDTA	ethylenediaminetetraacetic acid
ESD	environmental sex determination
FASW	filtered artificial sea water
FHA	forkhead-associated [domain]
Fox	forkhead box
GSD	genetic sex determination
HCR	hybridization chain reaction
HeSC	heteromorphic sex chromosome
HMM	hidden Markov model

HoSC	homomorphic sex chromosome
hpf	hours post fertilization
mRNA-ISH	mRNA *in-situ* hybridization
mab-3	*male abnormal-3*
ML	maximum likelihood
Mya	million years ago
ORF	open reading frame
PBS	1× phosphate-buffered saline
PBS-Tw	1× PBS with 0.1% Tween 20
PFA	paraformaldehyde
PGC	primordial germ cell
PSD	primary sex differentiation
RNAi	RNA interference
RT	room temperature
SC	sex chromosome
SDS	sodium dodecyl sulfate
SCO	single-copy orthogroup
SD	sex determination
SDG	sex-determining gene
Sox	*Sry*-related HMG-box
SRG	sex-determination related gene
Sry	*Sex-determining region of chromosome Y*
SSC-Tw	5× saline-sodium citrate with Tween 20 buffer
TBS	1× Tris-buffered saline
TBS-Tx	1× TBS with Triton X-100

Chapter 1

Introduction

1.1 The diversity of sexual processes

The process of sex determination (SD) has been traditionally associated with the very first step of gonad differentiation, where an initial trigger activates the molecular pathway that establishes organism sex. According to this view, two alternative types of SD have been recognized at first: the environmental sex determination (ESD) and the genetic sex determination (GSD), depending on whether the very first cues are of environmental or genetic origin. Conversely, all the downstream events of gonad development (i.e., after SD) have been appointed as primary sex differentiation (PSD), which consists of the entire set of morphogenetic, molecular, and physiological events leading to the full maturation of testes or ovaries (**Uller and Helanterä, 2011; Beukeboom and Perrin, 2014**). Lately, however, the dichotomous views of ESD/GSD and of SD/PSD have been questioned. On the one hand, a growing number of studies on non-model organisms proved that ESD and GSD represent a continuum of mixed conditions rather than two mutually exclusive phenomena. On the other, the high evolutionary dynamics and the variable expression patterns of the genes involved in the processes of gonad commitment and development make the distinction between SD and PSD of unclear utility (**Beukeboom and Perrin, 2014**).

Considering this complex scenario, **Uller and Helanterä, 2011** proposed a unified and broad-scope definition for SD, that is, "the processes within an embryo leading to the formation of differentiated gonads as either testes or ovaries", without any actual distinction between environmental/genetic initial triggers or the downstream effectors. However, I argue that this definition should be expanded to encompass not only the embryonic stage of the animal life cycle

but also adulthood, since cases of sex reversals and sex changes (sequential hermaphroditism) legitimately express proper SD processes during post-embryonic life stages as well.

1.2 Genetic sex determination and the evolution of sex-determining genes

In its most intimate core, animal SD is the manifestation of complex gene regulatory networks where, in accordance with the Wilkins' theory (1995), the downstream actors appear to be nearly conserved both from functional and identity point of views, while the master top regulators (the commonly recognized sex determinants, such as the *Sex-determining region of chromosome Y* (*Sry*) in therians or the ratio between sex and autosome chromosomes in *Drosophila*) are often the most variable part (**Beukeboom and Perrin, 2014**). As a matter of fact, this evolutionary pattern of animal sex-determining cascades has been observed in major animal clades, including vertebrates (e.g., **Marshall Graves and Peichel, 2010**), insects (e.g., **Verhulst et al., 2010**), and nematodes (e.g., **Stothard and Pilgrim, 2003**).

Sex-determination related genes (SRGs) are of particular interest not only from a regulatory point of view but also because of their patterns of molecular evolution. In fact, transcriptionally sex-biased genes (including SRGs) often tend to evolve faster than unbiased genes at the level of protein sequences. In particular, male-biased genes generally show higher rate of sequence evolution in comparison to both female-biased and unbiased counterparts (reviewed in **Parsch and Ellegren, 2013; Grath and Parsch, 2016**), as it has been repeatedly observed in well-studied organisms such as fruit flies (e.g., **Meisel and Connallon, 2013**), nematodes (e.g., **Cutter and Ward, 2005**), mice (e.g., **Kousathanas et al., 2014**) and primates (e.g., **Khaitovich et al., 2005**), and in other emerging model systems, such as *Daphnia pulex* (**Eads et al., 2007**), aphids (**Purandare et al., 2014**), and two wasp species of the genus *Nasonia* (**Wang et al., 2015**). Growing evidence is however showing cases in which instead female-biased genes have higher rates of sequence evolution than male-biased genes, such as in mosquitoes of the genus *Anopheles* (**Papa et al., 2017**), and European and Manila clams of the genus *Ruditapes* (**Ghiselli et al., 2018**).

The pattern of molecular evolution of sex-biased genes is particularly evident in organisms with sex chromosomes (both in XY/ZW and X0 systems), such as fruit flies, birds and mammals, where the so-called fast-X (or fast-Z) effect has been extensively reported for sex-chromosome

associated genes (**Vicoso and Charlesworth, 2006; Mank et al., 2007; Meisel and Connallon, 2013**). This high rate of sequence evolution in sex-biased genes and sex chromosomes (SCs) can be the result of both adaptative and non-adaptative processes, since the observed higher ratio between non-synonymous and synonymous mutations (dN/dS) can be caused by natural selection, sexual selection or sexual antagonism, as well as genetic drift (**Vicoso and Charlesworth, 2006; Meisel and Connallon, 2013; Parsch and Ellegren, 2013; Grath and Parsch, 2016**).

1.3 Sex determination in bivalves: a long-standing enigma

Bivalves are the second largest clade in molluscs, counting more than 18,000 species (Catalogue of Life) distributed at all depths and in all marine environments, as well as in some freshwater habitats. Thanks to their high diversity and biological peculiarities, they have been proposed as promising model organisms for investigating a wide array of biological, ecological and evolutionary issues (**Milani and Ghiselli, 2020; Ghiselli et al., 2021**). However, despite their socio-economic and scientific importance, the knowledge concerning the molecular basis of bivalve reproduction and SD is still quite limited (**Breton et al., 2018**). Clues from various works seem to suggest that both genetic and environmental factors (e.g., temperature, food availability, and steroids) are involved in SD, and that heteromorphic sex chromosomes (HeSCs) are absent (**Breton et al., 2018; Han et al., 2022**). However, the exact process by which sex is determined and gonad commitment is established is, currently, still unknown. Actually, bivalves represent a dazzling example of how the traditional dichotomies between ESD/GSD and SD/PSD can sometimes hamper scientific research, as many bivalve species exhibit various forms of hermaphroditism and because a master environmental or genetic sex determinant inducing PSD may just not exist.

In the attempt to identify SRGs, many differential gene expression analyses have been recently performed on a variety of species covering most of the phylogenetic diversity of bivalves (e.g., **Milani et al., 2013; Zhang et al., 2014; Chen et al., 2017; Capt et al., 2018; Ghiselli et al., 2018; Shi et al., 2018**). Some of the genes that were found to be differentially expressed between gonads of different sex were systematically retrieved across species, such as those belonging to the *doublesex* and *mab-3* related transcription factor (Dmrt), *Sry*-related HMG-box (Sox), and forkhead box (Fox) families, which act in concert in various animal

developmental processes including the SD cascade (**Marshall Graves and Peichel, 2010; Beukeboom and Perrin, 2014**). To this regard, **Zhang et al., 2014** proposed a working model for the sex-determining pathway of the Pacific oyster *Crassostrea gigas* in which: *CgSoxH* promotes male gonad development by activating *CgDsx*, which belong to the Dmrt family, and inhibiting *CgFoxL2*; *CgFoxL2*, when not inhibited by the pair *CgSoxH*/*CgDsx*, promotes female gonad development. Moreover, **Han et al., 2022** recently identified homomorphic sex chromosomes (HoSCs) in eight scallop species and appointed *FoxL2* as a putative SRG in *Patinopacten yessoensis* and *Chlamys farreri*. Though, much of the recent research effort on bivalve SRGs has been limited to their molecular cloning, differential transcription, and tissue localization (**Liang et al., 2019; Sun et al., 2022**). Furthermore, few works have directly investigated the biological functions of Dmrt, Sox, and Fox genes in bivalves so far, and most used post-transcriptional silencing of target mRNAs [RNA interference (RNAi)]. **Liang et al., 2019** studied the role of *Sox2* in the spermatogenesis of the Zhikong scallop *C. farreri* and found that it likely regulates proliferation of spermatogonia and apoptosis of spermatocytes, since its knockdown resulted in the loss of male germ cells. **Wang et al., 2020** proposed that in the female gonads of the freshwater mussel *Hyriopsis cumingii*, *FoxL2* might be related to the *Wnt*/*β-catenin* signaling pathway, which takes part in ovarian differentiation also in vertebrates. **Sun et al., 2022** found instead that in *C. gigas*, *FoxL2* and *Dmrt1L* mRNA knockdown results in the size reduction of female and male mature gonads, respectively.

In this sense, bivalve molluscs represent a striking example of the difficulty to reconcile the traditional view of a single sex determinant with an apparent multifactorial model in which many genes and environmental cues act in concert to establish the sexual identity of the individual (**Breton et al., 2018**). Lately, much effort has been put in the characterisation of bivalve SD and a general framework is eventually taking shape. Functional assays with RNAi and CRISPR-Cas9 techniques (e.g., **Wang et al., 2020; Sun et al., 2022; Wang et al., 2022**), as well as with mRNA *in-situ* hybridization (mRNA-ISH) and immunohistochemistry (e.g., **Perez-Garcia et al., 2011; Milani et al., 2013**), are making their way into the study of bivalve biology and have been proved essential instruments also for the investigation of sex-related traits. However, very few works have made extensive use of the comparative and integrative approach in bivalve studies so far, which hampers the possibility to infer general patterns for such a vast class of organisms (**Milani and Ghiselli, 2020**). The high evolutionary rates and plasticity of SRGs make the situation even harder, since phylogenetic and orthology

inferences can lead to erroneous reconstructions in the presence of signal saturation and high sequence divergence (reviewed in **Natsidis et al., 2021; Lozano-Fernandez, 2022**).

Chapter 2

Bivalves as emerging model systems to study the mechanisms and evolution of sex determination: a genomic point of view

Filippo Nicolini[1,2], Fabrizio Ghiselli[1], Andrea Luchetti[1], Liliana Milani[1]

[1] *Department of Biological, Geological and Environmental Science, University of Bologna, Bologna (BO), Italy.*

[2] *Fano Marine Center, Fano (PU), Italy.*

Published in: 2023, *Genome Biology and Evolution*, 15(10):evad181. 10.1093/gbe/evad181

Abstract. Bivalves are a diverse group of molluscs that have recently attained a central role in plenty of biological research fields, thanks to their peculiar life history traits. Here we propose that bivalves should be considered as emerging model systems also in sex-determination studies, since they would allow to investigate: (i) the transition between environmental and genetic sex determination, with respect to different reproductive backgrounds and sexual systems (from species with strict gonochorism to species with various forms of hermaphroditism); (ii) the genomic evolution of sex chromosomes, considering that no heteromorphic sex chromosomes are currently known and that homomorphic sex chromosomes have been identified just in few species of scallops; (iii) the putative role of mitochondria at some level of the sex determination signaling pathway, in a mechanism that may resemble the cytoplasmatic male sterility of plants; (iv) the evolutionary history of sex-determination related gene families with respect to other

animal groups. In particular, we think that this last topic may lay the foundations for expanding our understanding of bivalve sex determination, as our current knowledge is quite fragmented and limited to few species. As a matter of fact, tracing the phylogenetic history and diversity of sex-determination related gene families (such as the Dmrt, Sox and Fox genes) would allow to perform more targeted functional experiments and genomic analyses, but also fostering the possibility of establishing a solid comparative framework.

Significance. In this perspective, we provide an examination of the phylogenetic diversity of Dmrt genes, a sex-determination related gene family, to address the importance of bivalves in sex determination studies. By analyzing their taxonomic distribution and sequence diversity, we show how such a comparative study may set a common ground plan to settle down targeted functional experiments and essays. This kind of approach should be applied more extensively in future studies, especially when dealing with understudied organisms.

Bivalves are the second largest clade in molluscs, counting more than 18,000 species (Catalogue of Life, accessed 16/12/2022) distributed at all depths and in all marine environments, as well as in some freshwater habitats. Thanks to their high diversity and peculiar biological features, they have been proposed as promising model organisms for investigating a wide array of biological, ecological, and evolutionary issues, from mitochondrial biology and evolution to the physiological plasticity under fluctuating environmental conditions (**Milani and Ghiselli, 2020; Ghiselli et al., 2021**). In this context, bivalves may serve as a compelling model system to investigate the evolution and characteristics of sex determination (SD) as well, thanks to the diversity of their reproductive modes and genomic features. Nonetheless, this research field has been largely overlooked and many aspects of bivalve reproductive biology remain uncharacterized. In this perspective, we address the topic by first examining the relevant questions that bivalves may help to answer regarding processes and patterns of SD, and then providing a case study in the field of comparative genomics.

2.1 Open yet inspiring topics in bivalve sex determination

Despite the socio-economic and scientific importance of bivalves, the knowledge concerning the genetic and molecular bases of their SD system is quite limited and its study has been mostly neglected. Yet, bivalves may constitute a novel model system in SD studies that is as intriguing and valuable as other well-established models, such as vertebrates, insects and plants (**of Sex Consortium et al., 2014**), as they may provide complementary perspectives in many aspects of SD evolutionary studies. Topics such as (i) the transition between environmental and genetic SD, (ii) the evolution of sex chromosomes, (iii) the mito-nuclear interaction, and (iv) the evolution of SD related genes, can largely benefit from the integration with bivalve studies. But many others are likely to emerge as research in the field progresses.

2.1.1 Transitions between environmental and genetic sex determination

Clues from several works seem to suggest that both genetic and environmental factors are involved in bivalve SD, thus implying that a mixed system may exist (reviewed in **Breton et al., 2018**). The traditional dichotomy between environmental sex determination (ESD) and genetic

sex determination (GSD) seems inapplicable in most bivalve species, where ESD and GSD rather represent the two ends of a continuum of mixed and plastic conditions. A weak distinction between ESD and GSD is also found in amphibians, reptiles and teleost fish, three clades in which environment-dependent SD has been largely studied. Here, the interaction–or even the transition–between the two sexual systems have been reported in many species, suggesting that sex-determining mechanisms can be extraordinary plastic (**Bachtrog et al., 2014; Capel, 2017** Adding a representative and diverse group of Lophotrochozoa (Protostomia) to those vertebrate taxa, can widely expand the comparative framework of the investigation, allowing to better understand the evolution of SD as a whole. In bivalves, ESD has been studied mostly in oysters, where hermaphroditic species show an effect of temperature on SD (reviewed in **Breton et al., 2018**; **Fig. 2.1**). Oysters may indeed constitute a prolific model to examine how the SD pathways are shaped in the presence of different initial triggers and highly dynamic reproductive backgrounds. In fact, various sexual systems can be found in oysters, such as (i) strictly gonochoric population, (ii) the coexistence of simultaneous hermaphroditic with strictly gonochoric individuals in the same population, (iii) the possibility of sex change according to environmental conditions, and (iv) the presence of both parasitic dwarf males and free-living males in the same species (**Collin, 2013**). Consequently, oysters may be extremely useful to understand how epigenetic control is involved in sex change, how gene regulatory networks can sustain the occurrence of different hermaphroditic conditions within gonochoric populations, and whether certain SD systems are more labile then others (**Abbott, 2011**).

2.1.2 Evolution of sex chromosomes

So far, heteromorphic sex chromosomes (HeSCs)–i.e., sex chromosomes showing strong morphological differentiation, have never been observed in bivalves (**Breton et al., 2018**), while the first evidence of homomorphic sex chromosomes (HoSCs)–i.e., sex chromosomes showing little or no differentiation, comes from a very recent study on several scallop species, where a non-homologous origin of the SD system has been proposed for different subfamilies (**Han et al., 2022**; **Fig. 2.1**). Theory predicts that, once originated, sex chromosomes (SCs) will eventually turn into HeSCs, because of the recombination arrest in the sex-determining region (**Bachtrog et al., 2014; Beukeboom and Perrin, 2014; Han et al., 2022**). Nonetheless, HoSCs are much more widespread in the animal kingdom than expected, sometimes also being of ancient age (**Bachtrog et al., 2014; Han et al., 2022**).

Species from the order Pectinida may thus be useful to investigate what determines the long-term maintenance of HoSCs and which genomic architectures and molecular dynamics prevent HeSCs from evolving in bivalves. Additionally, they may be taken as model systems to investigate the origin of SCs in relation to the sexual systems and the route by which molecular pathways have been reprogrammed in the transition between different SD mechanisms (**Han et al., 2022**).

Researchers have been addressing this topic mainly in snakes, ratites and sturgeons (**Bachtrog et al., 2014; Han et al., 2022** and references therein). Though, scallops currently hold the oldest HoSC pairs, which are dated back to about 350 million years. The system is thus of great importance to investigate the role of sex-biased gene expression and selection forces in the long-term stability of SCs (**Han et al., 2022**), as well as the intertwining between SD systems.

2.1.3 Mito-nuclear interactions

An additional pivotal topic in bivalve biology, tentatively connected to SD, regards the doubly uniparental inheritance (DUI) of mitochondria, a process in which two highly divergent mitochondrial genomes are transmitted uniparentally through the maternal and paternal lineages, respectively through eggs and sperm. This process, which has been reported in more than a hundred bivalve species from five different orders (**Fig. 2.1**; **Gusman et al., 2016; Capt et al., 2020**), has been proposed to interact with the major nuclear pathways that primarily establish the sexual identity, in a way that can resemble the cytoplasmatic male sterility (CMS) of plants (**Ghiselli et al., 2013; Breton et al., 2022**). In CMS, specific mitochondrial chimeric open reading frames (ORFs) cause the pollen to be sterile, while certain nuclear loci act in counterbalance to restore male fertility when occurring in the same individual. This Red-Queen scenario, in which balancing selection shapes the evolution of both CMS and restorer-of-fertility genes and keeps the two sexes viable, has been also hypothesized to be acting on bivalve DUI species (**Ghiselli et al., 2013; Xu, Iannello, et al., 2022**), where additional and effectively-transcribed ORFs have been observed in both the male-inherited and female-inherited mitochondrial lineages (**Milani et al., 2013, 2014**).

Clearly, if a functional interplay between DUI and SD in bivalves is proven, this will provide new research questions regarding not only bivalve biology itself but also broader evolutionary topics (e.g., are there any converging trait between DUI and CMS systems? What is the degree

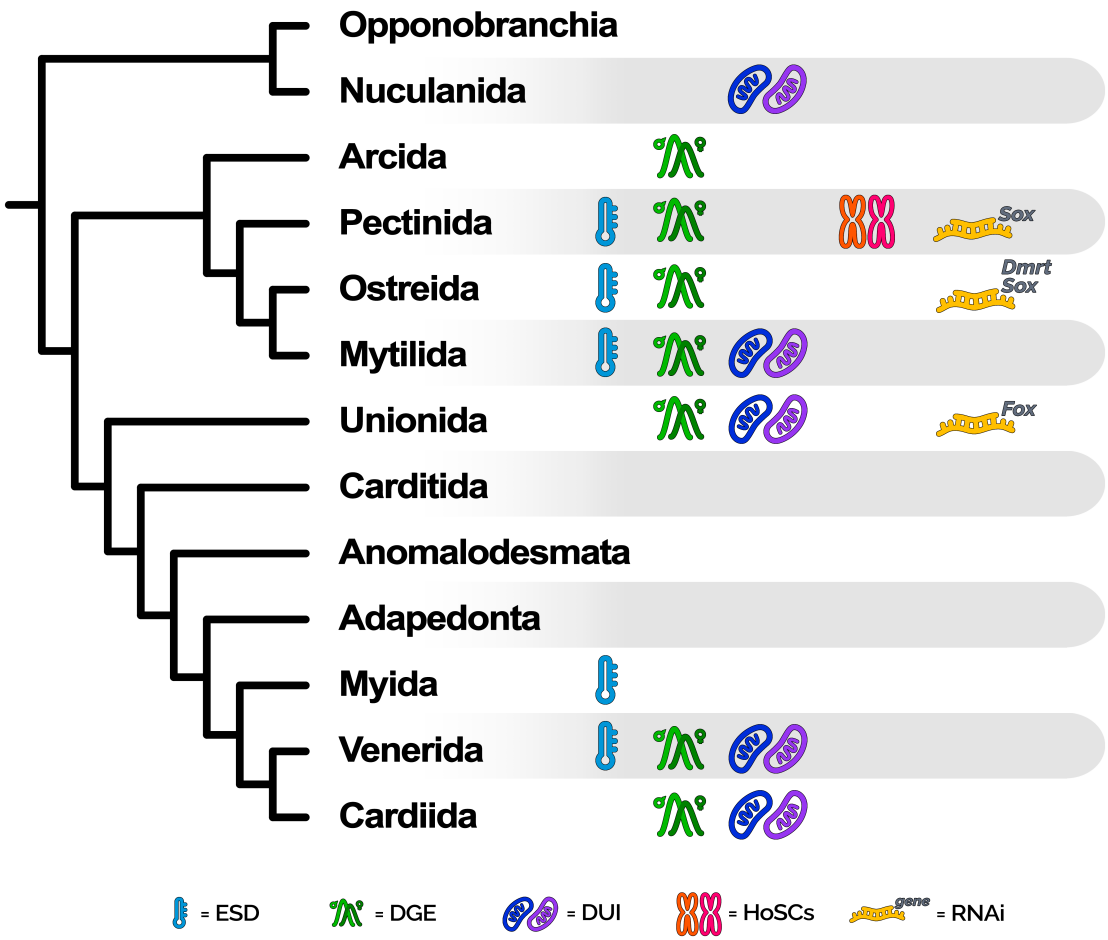


Figure 2.1. Graphical summary of the available knowledge and experiments concerning the genetic basis of SD in bivalves, at the level of major taxonomic orders (as reported in WoRMS; accessed before or on 14/03/2023). For each bivalve clade it is reported: (i) the availability of records of ESD (ii) the availability of differential gene expression (DGE) experiments specifically intended to investigate sex-biased or sex-specific genes; (iii) whether the DUI of mitochondria has been reported in at least one species; (iv) whether HoSCs have been identified in at least one species; (v) the availability of RNA interference (RNAi) experiments for genes belonging to the Dmrt, Sox, and Fox gene families. The phylogenetic tree on the left has been drawn on the basis of the most widely accepted topology for bivalves, according to analyses based on nuclear markers and morphological data. The tips of the tree correspond to major bivalve orders, except for Opponobranchia and Anomalodesmata, which represent higher-level taxonomic ranks. References for the availability of data and experiments can be found throughout the main test.

of plasticity of such mitochondria-related SD systems? Are mitochondria-related SD systems more widespread in eukaryotes than currently thought?).

2.1.4 Evolution of sex-determination related genes

Considering this intricate scenario of SD mechanisms and the wide diversity of bivalves, in the last years many differential transcription analyses have been performed on several species with the attempt to identify the most probable sex-determination related genes (SRGs) (e.g., **Milani et al., 2013; Zhang et al., 2014; Chen et al., 2017; Capt et al., 2018; Shi et al., 2018**; **Fig. 2.1**). Interestingly, certain genes consistently emerged across different bivalve species as being substantially more transcribed in one sex (sex-biased) or exclusively transcribed in one sex (sex-specific), suggesting their potential involvement in the SD pathway. These genes mainly belong to the *doublesex* and *mab-3* related transcription factor (Dmrt), *Sry*-related HMG-box (Sox), and forkhead box (Fox) families, which play a role in various developmental processes (including the SD cascade) in most animals (**Marshall Graves and Peichel, 2010; Bachtrog et al., 2014; Beukeboom and Perrin, 2014**). Members of these three gene families are also included in the working model for the SD regulatory network proposed for the Pacific oyster *Crassostrea gigas* by **Zhang et al., 2014**, in which: *CgSoxH* (which belong to the Sox family) promotes male gonad development by activating *CgDsx* (which belong to the Dmrt family) and inhibiting *CgFoxL2* (which belong to the Fox family); *CgFoxL2*, when not inhibited by the pair *CgSoxH*/*CgDsx*, promotes female gonad development. Similarly, **Han et al., 2022** appointed *FoxL2* as a putative SD gene in the two scallop species *Patinopacten yessoensis* and *Chlamys farreri*. If their pivotal role in SD of bivalves is confirmed, an evolutionary genomic analysis may help in better understanding why members of the above-mentioned gene families appear particularly prone to be recruited in the SD cascade also in distantly related species, as it is observed for *Dmrt1* and *Sox3* homologs in vertebrates (**Marshall Graves and Peichel, 2010; Bachtrog et al., 2014**; and the following section). Furthermore, considering the occurrence of mixed SD systems in bivalves, Dmrt, Sox, and Fox genes may provide new perspectives on the influence of different environmental cues on the molecular evolution of animal SRGs. However, to date, experiments have been limited to molecular cloning, differential transcription, and tissue localization of such genes (**Liang et al., 2019; Sun et al., 2022**), while only a few have directly investigated their biological functions in bivalves, for example through post-transcriptional silencing of target mRNAs [RNAi; **Fig. 2.1**; e.g., **Liang et al., 2019; Wang**

et al., 2020; Sun et al., 2022].

Overall, Dmrt, Sox, and Fox genes are highly interesting targets to be investigated in the framework of bivalve SD and have indeed obtained much more attention than the study of SCs or the role of environmental cues. However, much work is still to be done in order to understand their function in the SD signaling pathway and their evolutionary history.

2.2 The case of the Dmrt gene family in bivalves

Among the SRG candidates identified in bivalves, Dmrt genes (named after *doublesex* (*dsx*) from *Drosophila melanogaster* and *male abnormal-3* (*mab-3*) from *Caenorhabditis elegans*) are of particular interest. As a matter of fact, in vertebrates, besides their role in placode neurogenesis and somite patterning (reviewed in **Mawaribuchi et al., 2019**), Dmrt genes are also involved in the development of male gonads and the maintenance of the testicular function (**Sun et al., 2022**). Their role in the specification and organization of male sexual characters seems indeed to be common across Metazoa, suggesting that a similar function may have been already present in the Bilateria common ancestor (**Kopp, 2012; Beukeboom and Perrin, 2014**).

The first attempts to dig inside the phylogenetic history and diversity of bivalve Dmrt genes have been provided by **Li et al., 2018** and **Evensen et al., 2022**: besides retrieving all the canonical genes (i.e., *Dmrt2*, *Dmrt3* and *Dmrt4/5*), their inferences brought to light a monophyletic Dmrt group (named *doublesex and mab-3 related transcription factor 1-like* (*Dmrt-1L*)) which appears to be private to molluscs and present in several bivalve species. The *Dmrt-1L* monophyletic group is confirmed also when expanding the analysis by mining genomes from a wider range of bivalve taxa (**Fig. 2.1**; **Fig. 2.2A**), suggesting that *Dmrt-1L* genes are widespread in bivalves and were likely present in their common ancestor (**Evensen et al., 2022**). In particular, *Dmrt-1L* genes can been successfully retrieved in species of the orders Mytilida, Ostreida, Pectinida, Unionida, and from *Scapharca broughtonii* (Arcida), while the opposite holds for Venerida, *Sinonovacula constricta* (Adapedonta), and *Dreissena* spp. (Myida; **Fig. 2.2B**). Clearly, the absence of *Dmrt-1L* genes demands further investigations, as it may derive from errors in genome assembly and annotations.

The present analysis also supports a higher amino acid sequence divergence of the *Dmrt-1L* orthology group with respect to the other Dmrt orthology groups (**Fig. 2.1C**), which may

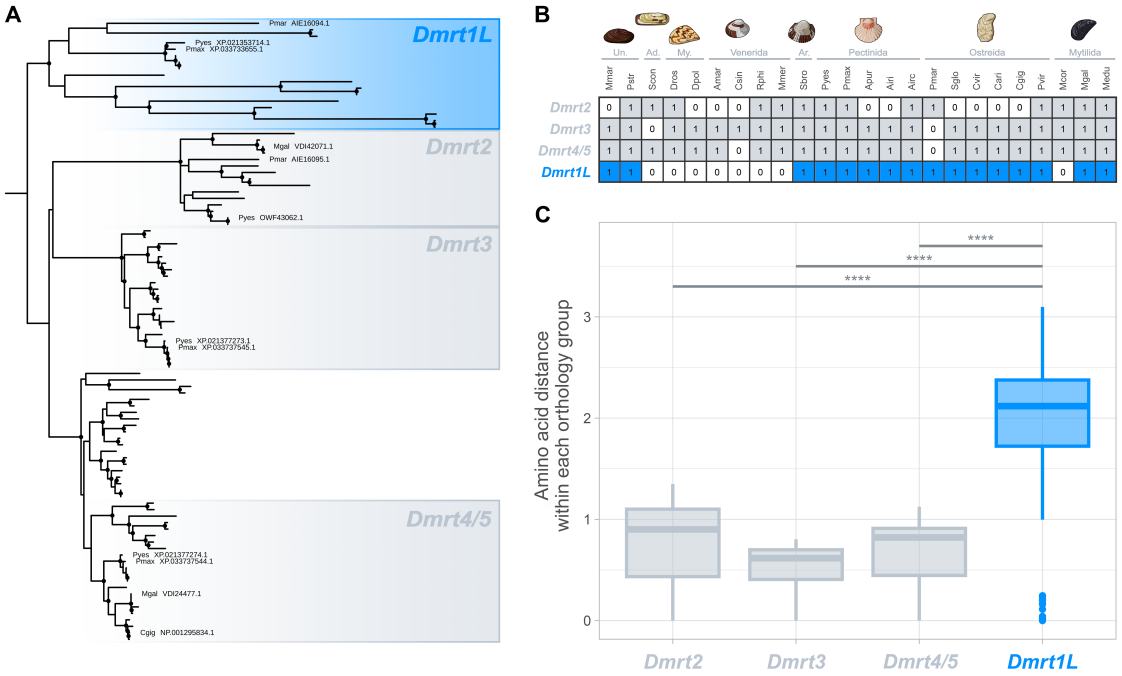


Figure 2.2. Phylogenetic tree (A) and taxonomic distribution (B) of Dmrt genes in bivalves, and comparison of amino acid pairwise distances within *Dmrt-1L* and the other Dmrts (B). (A) Dmrt orthologs from bivalve genome assemblies were obtained with HMMsearch (HMMER toolkit; **Eddy, 2011**) with the Pfam HMM profile of the DM domain (PF00751). Amino acid alignment was obtained with MAFFT-DASH (**Rozewicki et al., 2019**), and manually inspected to remove poorly aligning sequences, and trimmed with trimAl (gap threshold of 60%; **Capella-Gutiérrez et al., 2009**). The phylogenetic analysis was carried out using IQ-TREE 2 (**Minh et al., 2020**) with default parameters. Nodes with bootstrap values greater than 84 are marked with filled black circles. The tree was rooted according to **Evensen et al., 2022**. Dmrt genes analysed by **Evensen et al., 2022** were used as reference to annotate the various orthology groups, and accession numbers are reported in the tree. The phylogenetic tree with all annotated tips and nodes can be accessed on supplementary material online. (B) Taxonomic distribution of identified Dmrt genes in bivalve genomes. Orders as reported in WoRMS (accessed before or on 14/03/2023) and in **Fig. 2.1** are specified. (C) Pairwise amino acid distances were computed for amino acid sequences within each Dmrt orthology group identified in the tree, with the R package 'phangorn' (**Schliep, 2011**) under the JTT substitution model. After checking for normality with the Shapiro-Wilk test (W = 0.88544, p-value ¡ 2.2e-16) and for group effect with the Kruskal-Wallis test (p-value ¡ 2.2e-16), the pairwise Wilcoxon rank-sum test was used to compare the distributions of pairwise amino acid distances of Dmrt-1L and the other Dmrts. Horizontal bars mark the significative results with p ¡ 2.2e-16 ('****') (Bonferroni correction for multiple test was applied). The list of genome assemblies used for these analyses and species identifiers can be found in **Fig. 2.1**. Un.: Unionida; Ad.: Adapedonta; My.: Myida; Ar.: Arcida.

be explained by a higher rate of sequence evolution related to their sex-biased expression in certain species (**Zhang et al., 2014; Shi et al., 2015; Li et al., 2018; Evensen et al., 2022**). This is consistent with what has been already observed for the SRGs *Dmrt1* and *dsx* in vertebrates and *Drosophila*, respectively (e.g., **Bewick et al., 2011; Baral et al., 2019**). In fact, sex-biased genes (including SRGs) often tend to evolve faster than unbiased genes at the level of protein sequences, either when considering male-biased (reviewed in **Parsch and Ellegren, 2013; Grath and Parsch, 2016**) or female-biased genes (e.g., **Papa et al., 2017; Ghiselli et al., 2018**). Another possible explanation for the higher amino acid divergence of *Dmrt-1L* genes may lie on their expression breadth, that is, genes with a narrow tissue-specific expression tend to evolve faster than more ubiquitous genes (**Parsch and Ellegren, 2013; Xu, Martelossi, et al., 2022**). As a matter of fact, *Dmrt-1L* genes have been found to be significantly more transcribed in the gonadic tissue (particularly in testes) in *P. yessoensis* (**Li et al., 2018**) and *C. gigas* (**Yue et al., 2021**).

Understanding the role and molecular interactions of *Dmrt-1L* genes in bivalve SD and gonad development would greatly enhance the possibility of outlining the evolutionary causes and consequences of their high amino acid divergence (**Fig. 2.2C**), for example by linking the molecular evolution to the degree of pleiotropy. However, most of our knowledge on *Dmrt-1L* biology is currently limited to the temporal and tissue localization of transcripts in a few species of bivalves (e.g., **Li et al., 2018; Yue et al., 2021**). In fact—apart from the work by **Sun et al., 2022**, which confirmed the role of *Dmrt-1L* in the gonad development of *C. gigas* through non-invasive RNAi and found that the knocked-down phenotype results in size reduction of male gonads—no other experiments intended to elucidate the function of *Dmrt-1L* genes in bivalves have been carried out so far (**Fig. 2.1**). This clearly hinders any possible integration between molecular data with functional assays. If the role of *Dmrt-1L* as major sex determinants was confirmed, bivalves would become an intriguing clade in which investigate why, in Metazoa, certain genes (namely, the Dmrt gene family) appear particularly prone to being recruited at the top of the SD cascade. To date, this phenomenon has been widely examined in vertebrates, where *Dmrt1* genes have independently gained a primary role in male SD in fish, amphibians, and birds, and are considered candidate sex-determining genes also in monotreme mammals (**Marshall Graves and Peichel, 2010; Beukeboom and Perrin, 2014; Mawaribuchi et al., 2019**). Bivalves may provide an alternative evolutionary scenario to study the selective forces and molecular modifications that support Dmrt genes in repeatedly taking over the SD

process. In fact, since *Dmrt-1L* genes seem to be restricted to molluscs (**Fig. 2.2A**), it would be intriguing to clarify if the putative involvement in the SD cascade of extant bivalve species is the result of shared ancestry or convergent evolution, which would establish a study system for the evolution of Dmrt genes parallel to that of vertebrates (see **Capel, 2017**).

Obviously, *Dmrt-1L* should not be expected to be the sole sex-determining gene. In fact, *Fox-L2* has already been appointed as the female sex-determining gene in *P. yessoensis* and *C. farreri* (**Han et al., 2022**). Consequently, we should expect that other primary genetic determinants exist, consistently with the extremely high species diversity of the clade. Thus, bivalves may additionally serve as a valuable model system to study how genes from different families take over the SD cascade and are shaped by selection.

2.3 Conclusions: bivalves as new models in the study of sex determination

SD is undoubtedly a fascinating biological and evolutionary topic as much as it is challenging to investigate. Our understanding of the causes and consequences of the SD mechanism diversity strongly relies on the study of different systems and non-model model organisms (**Bachtrog et al., 2014; Milani and Ghiselli, 2020**), which provide the foundation for depicting a comprehensive evolutionary and comparative framework in which new and coherent research perspectives can be grounded.

In recent years, bivalves have been achieving growing importance in many fields of biology, from ecology to genomics, and from environmental biomonitoring to mitochondrial studies (**Milani and Ghiselli, 2020; Ghiselli et al., 2021**), but they can be a valuable model to address also SD studies. The diversity of their life history traits provides indeed a challenging, yet extremely fascinating framework, to put the SD processes into an evolutionary context.

Bivalves can help us explain how ESD and GSD interplay with each other in response to the environmental conditions, as a mixed system of both has been proposed to act in the establishment of bivalve sexual identity (reviewed in **Breton et al., 2018**). Moreover, the occurrence of the many existing variants of hermaphroditism and gonochorism even in closely related species, or within the same population, strongly suggests that the basic SD pathway (whether genetic, environmental, or mixed) should be plastic enough to sustain the existence

of individuals of both sexes, thus providing the opportunity to study how SD gene regulatory networks are shaped and selected throughout evolution and how epigenetic regulation may influence SD. The unique DUI system further poses an undeniable challenge in SD studies since it may represent an SD-linked mechanism which relies on the non-nuclear portion of the genome and may unfold many new research paths (**Milani and Ghiselli, 2020; Ghiselli et al., 2021**). Nonetheless, much of the research effort on bivalve SD has been devolved to specific groups of socio-economic importance, such as Mytilida, Ostreida, Pectinida, and Unionida, while the other lineages of the bivalve phylogeny have been neglected (**Fig. 2.1**). Our understanding of the SD processes of bivalves is thus restricted and is mainly lacking a broad comparative framework in which to draw comprehensive evolutionary inferences.

Genes from the Dmrt, Sox and Fox families, which are involved in SD also in other Metazoa, may be considered excellent genomic targets to study the processes and patterns of molecular evolution in sex-biased genes, as well as of the recurrent recruitment of genes in the SD cascade. Also, identifying the major genetic regulators of SD in bivalves would burst the functional study of the interaction between ESD and GSD, by providing genetic targets that can be manipulated through RNAi and/or genome editing techniques to understand the role of environmental cues in SD. In the same way, knowing the main genetic actors of SD would allow researcher to identify SCs not only on the basis of in-silico techniques (such as k-mer based or SNP methods) but also by less-expensive wet lab protocols (such as fluorescence mRNA *in-situ* hybridization (mRNA-ISH) on metaphase chromosome plates). Furthermore, it would help to understand whether and how the mitochondrial additional ORFs of DUI species interact with the SD system, by performing thorough gene expression essays.

In conclusion, we strongly urge researchers to invest more resources in the integrative study of bivalve SD to unravel the many underlying mechanisms and expand our understanding of this biological process. Given our limited knowledge in the field, one of the first routes that should be undertaken may rely on the comparative study of SRGs of bivalves from a genomic perspective, as this kind of data is nowadays growing at a rate faster than ever. Establishing such a genomic ground plan for understudied organisms will in fact allow researchers to develop evolutionary-aware experiments with better selected genetic targets.

Table 2.1. List of bivalve genomes from which Dmrt genes have been extracted. For each species, the accepted name and the most-common synonym (in parentheses) are reported. NCBI accession numbers are provided, when available, as well as BUSCO scores of the predicted proteomes against the metazoa_odb10 dataset (**Manni et al., 2021**).

Species	**ID**	**Order**	**Assembly level**	**BUSCO score**	**Reference**	**NCBI Acc. No.**
Anadara (Scapharca) broughtonii	Sbro	Arcida	Chromosome	C:91.2% [S:85.6%,D:5.6%] F:2.6% M:6.2%	**Bai et al., 2019**	NA
Sinonovacula constricta	Scon	Adapedonta	Chromosome	C:92.5% [S:80.4%,D:12.1%] F:3.4% M:4.1%	**Ran et al., 2019**	GCA_007844125.1
Dreissena polymorpha	Dpol	Myida	Chromosome	C:86.9% [S:75.1%,D:11.8%] F:6.4% M:6.7%	**McCartney et al., 2022**	GCA_020536995.1
Dreissena rostriformis	Dros	Myida	Scaffold	C:75.2% [S:73.2%,D:2.0%] F:15.2% M:9.6%	**Calcino et al., 2019**	GCA_007657795.1
Mytilus unguiculatus (coruscus)	Mcor	Mytilida	Chromosome	C:80.0% [S:79.1%,D:0.9%] F:7.7% M:12.3%	**Yang et al., 2021**	GCA_017311375.1

Tab. 2.1 continued from previous page

Species	ID	Order	Assembly level	BUSCO score	Reference	NCBI Acc. No.
Mytilus edulis	Medu	Mytilida	Scaffold	C:83.7% [S:64.5%,D:19.2%] F:5.2% M:11.1%	**Corrochano-Fraile et al., 2022**	GCA_905397895.1
Mytilus galloprovincialis	Mgal	Mytilida	Scaffold	C:80.3% [S:47.5%,D:32.8%] F:8.8% M:10.9%	**Gerdol et al., 2020**	GCA_900618805.1
Perna viridis	Pvir	Mytilida	Scaffold	C:99.4% [S:99.0%,D:0.4%] F:0.2% M:0.4%	**Inoue et al., 2021**	GCA_018327765.1
Magallana (Crassostrea) ariakensis	Cari	Ostreida	Chromosome	C:94.6% [S:90.9%,D:3.7%] F:0.9% M:4.5%	**Li et al., 2021**	GCA_020567875.1
Magallana (Crassostrea) gigas	Cgig	Ostreida	Chromosome	C:98.5% [S:67.6%,D:30.9%] F:0.3% M:1.2%	**Peñaloza et al., 2021**	GCF_902806645.1

Tab. 2.1 continued from previous page

Species	ID	Order	Assembly level	BUSCO score	Reference	NCBI Acc. No.
Crassostrea virginica	Cvir	Ostreida	Chromosome	C:98.1% [S:58.6%,D:39.5%] F:0.3% M:1.6%	**Gómez-Chiarri et al., 2015**	GCF_002022765.2
Saccostrea glomerata	Sglo	Ostreida	Scaffold	C:88.9% [S:85.3%,D:3.6%] F:5.1% M:6.0%	**Powell et al., 2018**	GCA_003671525.1
Argopecten irradians concentricus	Airc	Pectinida	Scaffold	C:94.8% [S:93.9%,D:0.9%] F:3.7% M:1.5%	**Liu et al., 2020**	GCA_004382765.1
Argopecten irradians irradians	Airi	Pectinida	Scaffold	C:94.8% [S:93.9%,D:0.9%] F:3.7% M:1.5%	Liu et al., 2020	GCA_004382745.1
Argopecten purpuratus	Apur	Pectinida	Scaffold	C:89.2% [S:88.5%,D:0.7%] F:5.0% M:5.8%	**Liu et al., 2020** NA	

Tab. 2.1 continued from previous page

Species	ID	Order	Assembly level	BUSCO score	Reference	NCBI Acc. No.
Pecten maximus	Pmax	Pectinida	Chromosome	C:98.5% [S:74.7%,D:23.8%] F:0.4% M:1.1%	**Kenny et al., 2020**	GCF_902652985.1
Mizuhopecten (Patinopecten) yessoensis	Pyes	Pectinida	Scaffold	C:98.6% [S:75.2%,D:23.4%] F:0.4% M:1.0%	**Wang et al., 2017**	GCF_002113885.1
Margaritifera margaritifera	Mmar	Unionida	Scaffold	C:92.6% S:82.3%,D:10.3%] F:3.2% M:4.2%	**Gomes-dos-Santos et al., 2021**	GCA_015947965.1
Potamilus streckersoni	Pstr	Unionida	Scaffold	C:74.7% [S:73.8%,D:0.9%] F:7.0% M:18.3%	**Smith, 2021**	GCA_016746295.1
Calyptogena (Archivesica) marissinica	Amar	Venerida	Chromosome	C:82.0% [S:80.0%,D:2.0%] F:6.1% M:11.9%	**Ip et al., 2021**	GCA_014843695.1

Tab. 2.1 continued from previous page

Species	ID	Order	Assembly level	BUSCO score	Reference	NCBI Acc. No.
Cyclina sinensis	Csin	Venerida	Scaffold	C:94.0% [S:83.8%,D:10.2%] F:1.9% M:4.1%	**Wei et al., 2020**	GCA_012932295.1
Mercenaria mercenaria	Mmer	Venerida	Chromosome	C:95.4% [S:70.9%,D:24.5%] F:0.5% M:4.1%	**Song et al., 2021**	GCF_014805675.1
Ruditapes philippinarum	Rphi	Venerida	Chromosome	C:83.4% [S:74.5%,D:8.9%] F:8.8% M:7.8%	**Xu, Martelossi, et al., 2022**	GCA_026571515.1

2.4 Acknowledgments

The authors are extremely thankful to Sofía Blanco González from the University of Vigo for her willingness to engage in discussions and for genuinely sharing her opinion on this work.

2.5 Data Availability

Analyzed data and R scripts used to generate plots can be accessed in supplementary material online deposited at the following GitHub repository: filonico/bivalve_sex_perspective.

Chapter 3

Identification of putative sex-determination related genes in bivalves through comparative molecular evolutionary analyses

Filippo Nicolini[1,2], Mariangela Iannello[1], Giovanni Piccinini[1], Sergey Nuzhdin[3], Fabrizio Ghiselli[1], Andrea Luchetti[1], Liliana Milani[1]

[1]*Department of Biological, Geological and Environmental Science, University of Bologna, Bologna (BO), Italy.*

[2]*Fano Marine Center, Fano (PU), Italy.*

[3]*Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA.*

In preparation.

3.1 Introduction

In preparation.

3.2 Materials and Methods

3.2.1 Dataset of bivalve annotated genomes and transcriptomes

Annotated genome assemblies of bivalves were obtained from various publicly available resources, while reference genome assemblies for gastropods and cephalopods were downloaded from NCBI (**Supp. Tab. S3.1**). Isoforms were removed from genome annotations using a perl script from the AGAT toolkit (v0.8.0; **Dainat, unpublished**). Concerning *Sinonovacula constrica* (Adapedonta), the nucleotide coding sequence fasta file was not available for download. To avoid excluding the species from our analyses, the file was generated in-house by mapping the annotated protein sequences on the reference genome using miniprot (v0.13-0; **Li, 2023**). Then, the corresponding nucleotide sequences were extracted using AGAT on the resulting gff annotation file.

In order to provide an extensive identification of sex-determination related genes (SRGs) also for underrepresented bivalve orders (mainly belonging to the Heterodonta clade), 14 additional species represented by sequenced transcriptomes were included in the analyses. Transcriptomes were obtained, assembled and annotated following **Piccinini et al., 2021** and **Iannello et al., 2023**. Briefly, raw reads were trimmed using Trimmomatic (**Bolger et al., 2014**) and assembled using Trinity (**Grabherr et al., 2011**) with default parameters. Isoforms were removed using the dedicated perl script from the Trinity utilities. Open reading frames were predicted through TransDecoder (**Haas, unpublished**; github.com/TransDecoder), by also including diamond (**Buchfink et al., 2015**) and HMMER (v3.3.2; hmmer.org) annotation of hits.

The resulting set of annotated genomes and transcriptomes (hereafter referred to as the 'comprehensive set') was checked for completeness using BUSCO with the Metazoa reference dataset (v5.2.2; **Manni et al., 2021**).

3.2.2 Identification and classification of Dmrt, Sox and Fox genes in bivalves

Members of the Dmrt, Sox and Fox gene (DSFG) families were retrieved in the comprehensive set with hmmsearch from the HMMER package (v3.3.2; hmmer.org). The signature catalytic domains of each family were used as queries. Specifically, HMM profiles were built after the Pfam databases for the *dsx* and *mab-3* (DM) domain (PF00751), the high mobility group (MHG) box (PF00505) and the forkhead domain (PF00250) to retrieve members of the DSFG families, respectively. The e-value for both the per-target and the per-domain inclusion threshold was set to 10^{-5}.

Obtained hits were then annotated using (i) the PANTHER HMM standalone sequence scoring against the PANTHER library v18.0 and (ii) RPS-BLAST (v2.5.0+) against the Conserved Domain Database (CDD; pre-compiled version, downloaded from ftp.ncbi.nih.gov on 09/11/23). In both cases, hits with an e-value of 10^{-5} were retained. Genes which were correctly annotated by both systems (on the basis of the PANTHER gene family and CDD domain identifiers; **Supp. Tab. S3.2**) were kept for subsequent analyses. Dmrt, Sox, and Fox genes from *Homo sapiens*, *Drosophila melanogaster*, and *Caenorhabditis elegans* (**Supp. Tab. S3.3**; hereafter referred to as 'reference species') were retrieved from NCBI and were used as reference genes for annotation (see below). Classification and nomenclature of each family was retrieved from: **Mawaribuchi et al., 2019** for Dmrt genes; **Phochanukul and Russell, 2010** and **Sarkar and Hochedlinger, 2013** for Sox genes; **Mazet et al., 2003** for Fox genes.

The alignments of mollusc and reference DSFGs were guided by the aforementioned Pfam HMM profiles and performed with Clustal Omega (v1.2.3; **Sievers et al., 2011**), then trimmed with trimAl (v1.4.rev15; **Capella-Gutiérrez et al., 2009**) with a gap threshold of 40%. Resulting alignments were manually inspected to remove sequences with incomplete catalytic domains, then aligned and trimmed again as before. Phylogenetic trees were inferred using IQ-TREE (v2.1.4-beta COVID-edition; **Minh et al., 2020**) with automatic model selection (**Kalyaanamoorthy et al., 2017**), 1000 bootstrap replicates and 5 independent runs. The phylogenetic tree of Dmrt genes was midpoint rooted, as no clear homology relationship has been found with other gene families or zinc-finger proteins so far (**Wexler et al., 2014**). Phylogenetic trees of Sox and Fox gene families were rooted using two fungi MATA-1 sequences (XP_62685912.1, CCD57795.1) and two Amoebozoa forkhead-like domains (XP_004368148.1,

XP_004333268.1), respectively (**Nakagawa et al., 2013; Heenan et al., 2016**). The rooting was performed with Gotree (v0.4.5; **Lemoine and Gascuel, 2021**). To identify and annotate bivalve homology groups within each gene family, we employed a species overlap algorithm followed by an MCL clustering weighted by node supports as implemented in Possvm (v1.2; **Grau-Bové and Sebé-Pedrós, 2021**). DSFGs from *H. sapiens*, *D. melanogaster*, and *C. elegans* were used as reference annotation.

In order to better establish the orthology relationships among ambiguous groups of Dmrt and Fox genes, we run a series of other phylogenetic reconstructions (see **Discussion**), by using the same pipeline as before. In the case of *Fox-Y* genes, we also employed Fox gene sequences from the sea urchin *Strongylocentrotus purpuratus*, as given by **Tu et al., 2006**. All the phylogenetic trees were plotted using the R package 'ggtree' (**Yu, Smith, et al., 2017**).

3.2.3 Sequence diversity of bivalve single-copy orthogroups

As a metrics to measure the sequence diversity of bivalve DSFGs, and test whether those putatively involved in sex determination (SD) show higher values than other genes, we employed the amino acid sequence divergence (AASD). As a matter of fact, this metric is pretty fast and straightforward to obtain, as it only requires the amino acid alignment and the corresponding best-fit substitution mode.

To this purpose, we produced amino acid alignments of bivalve single-copy orthologous (SCOs) groups and built the distribution of their median AASD. Specifically, we assembled a second dataset (hereafter referred to as the 'reduced bivalve dataset') which includes, for each bivalve genus, only the best genomes and transcriptomes in terms of either BUSCO scores (on the metazoan_odb10 dataset) or assembly statistics (**Supp. Tab. S3.1**), in order to reduce computational time. *Archivesica marissinica* and *Saccostrea glomerata* were also removed, as their annotated coding sequences contain many stop codons, which prevent accurate amino acid guided alignments. Genes were clustered in orthologous groups using OrthoFinder (v2.5.5; **Emms and Kelly, 2019**) with DIAMOND ultra-sensitive and default parameters. Resulting orthogroups were splitted into SCOs using DISCO (v1.3.1; **Willson et al., 2022**), and orthogroups with at least 17 species (50% of the species included in the bivalve reduced dataset) were retained. Amino acid and nucleotide sequences of SCOs were then aligned using Clustal Omega as implemented in TranslatorX (v1.1; **Abascal et al., 2010**), and jointly trimmed using trimAl with a gap threshold of 40% and the removal of spurious sequences (`-resoerlap`

`50 -seqoverlap 50`). Eventually, orthogroups containing (i) internal stop codons, (ii) with less than 17 species left (50% of the species included in the bivalve reduced dataset), or (iii) containing DSFGs were removed from downstream analyses. The best amino acid substitution model was inferred for each trimmed alignment using ModelFinder as implemented in IQTREE2 (model search was restricted to matrices accepted by the 'phangorn' R library, i.e., Blosum62, cpREV, Dayhoff, DCMut, FLU, HIVb, HIVw, JTT, JTTDCMut, LG, mtART, mtMAM, mtREV, mtZOA, rtREV, VT, WAG) and the corresponding pairwise amino acid distances were computed with the function 'dist.ml' from the 'phangorn' R package (**Schliep, 2011**). We decided to employ the pairwise amino acid distance instead of the tip-to-tip phylogenetic distance in order to save computational time. However, to check whether the two metrics were comparable to each other, we randomly selected 200 decomposed orthogroups (including orthogroups from the DSFGs) and computed the maximum likelihood (ML) trees using IQTREE2, with ModelSelection restricted as before. Then, the tip-to-tip pairwise distances were obtained with the R package 'adephylo' (**Jombart and Dray, 2010**). The same pipeline was also employed to obtain pairwise amino acid distances for each DSFG single-copy orthologous group.

The distribution of amino acid distances was then built after the median values of pairwise distances of each SCO, and genes were categorised accordingly into three groups: Group 1, consisting of genes from the 1% upper quantile of the distribution; Group 2, consisting of genes between the 1% and 5% upper quantiles; and Group 3, consisting of all the remaining genes.

3.2.4 Mammals and *Drosophila* spp. as test datasets

To validate our approach for the study of bivalve SRG molecular evolution, we run the same analysis on two additional datasets, consisting of reference genomes of mammals and *Drosophila* species (**Supp. Tab. S3.4–S5**, respectively), whose SD process is well studied and characterised. As a matter of fact, despite it is well known that sex-determining genes (SDGs) tend to evolve faster than genes not involved in SD, the hypothesis has never been tested extensively across the entire phylogenetic diversity of a group: to the best of our knowledge, molecular evolution of SDGs and SRGs has always been tested on single species or inside the boundaries of taxonomic genera (REFERENCE). We chose mammals and *Drosophila* as they provide different frameworks to study the patterns of molecular evolution in SDGs: the former is a system where SD is completely genetic (where the development into a male or into a female is

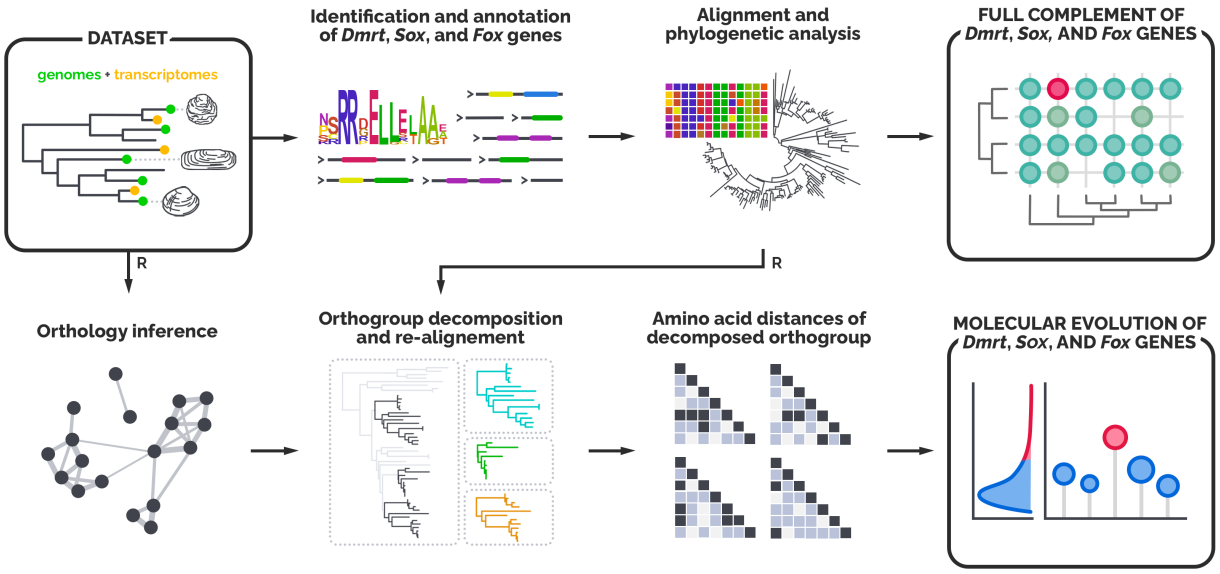


Figure 3.1. Workflow of the analyses for the bivalve dataset. Starting from a set of both genomes and transcriptomes covering a great portion of bivalve taxonomic diversity, we first characterized the entire complement of glsdsfg genes (upper row). In particular, we used sequence annotation and phylogenetic tools to obtain reliable sequences and filter out any putative mis-assembled or mis-annotated sequence. Afterwards, we built a reduced set of transcriptomes and genomes (the reduced bivalve dataset, where we minimized the redundancy of congeneric species) from which to draw the molecular evolution patterns of orthologous genes (bottom row). In particular, after having obtained gene single-copy orthologous groups, we calculated the amino acid distances within each orthogroup and then we built the distribution of median values. The same pipeline was also employed for the mammal and the fruit fly datasets, with just two minor differences: the starting dataset was composed of only genomes, and that the reduction step (R) was not necessary.

triggered by the up- or downregulation of *Sry* in undifferentiated gonads, respectively), while the latter is a system where SD is chromosomic, thus lacks a master SDG (the sexual fate of the individual is determined by the ratio between autosomal and X chromosomes). Consequently, they represent opposing (and possibly overlapping) control datasets to be compared to bivalves. For both mammals and fruit flies, annotated genomes were downloaded from NCBI using the command-line tool 'datasets', then processed using the same pipeline and scripts as before (**Fig. 3.1**).

3.2.5 GO-term enrichment

After having obtained the distributions of AASD in the three datasets (Bivalvia, Mammalia, and *Drosophila*) and having sorted SCOs genes up into 3 groups (Group 1, group 2, and group 3), we performed a gene ontology (GO) enrichment analysis of genes from Group 1 and genes from Group 1 + Group 2. To do so, we firstly selected one gene per SCO, giving priority to few chosen

species: (i) for bivalves, we selected genes from *P. maximus*, or alternatively from *C. gigas*, *Hyriopsis bialata*, *Tridacna squamosa*, and *Solen grandis*; (ii) for mammals, we selected genes from *H. sapiens*, or alternatively from *Bubalus bubalis*, *Panthera tigris*, *Camelus dromedarius*, and *Monodelphis domestica*; (iii) for fruit flies, we selected genes from *D. melanogaster*, or alternatively from *Drosophila ananassae*, *Drosophila hydei*, and *Drosophila suzukii*. By doing so, we ensured that each SCO was represented by one gene. Afterwards, we annotated the obtained datasets with the corresponding GO terms using the OMA browser (accessed 18/09/2024; **Altenhoff et al., 2024**). The GO-term enrichment of Group 1 genes and Group 1 + Group 2 genes was performed with the R package 'topGO', using the Fisher exact test (**Alexa and Rahnenführer, 2009**).

3.3 Results

3.3.1 Genomic and transcriptomic datasets

The complete bivalve dataset consists of 29 bivalve genomes, 14 bivalve transcriptomes, and 7 outgroup genomes (5 gastropods and 2 *Octopus* spp.; **Supp. Tab. S3.1**). BUSCO statistics for complete single-copy genes spanned from the 64.9% in *Modiolus modiolus* to the 99.4% of *Perna viridis*, with a median value of 94.7%. We were able to get at least one representative species for 11 different bivalve orders, covering a good proportion of the phylogenetic diversity of the clades Pteriomorphia, Palaeoheterodonta, and Imparidentia, and thus building the most extensive genomic and transcriptomic dataset for bivalve comparative analyses so far (**Supp. Tab. S3.1**). Unfortunately, no genomes or transcriptomes for Protobranchia, Archiheterodonta, and Anomalodesmata were available at the time of the project, thus we were not able to include any of those clades in our analysis. The reduced bivalve dataset (used for the orthology inference and the molecular evolution analysis; **Fig. 3.1**) consists instead of 36 genomes and transcriptomes (**Supp. Tab. S3.1**), and was built to retain just one species for each taxonomic genera.

The mammal dataset consists of 32 species and 1 outgroup (*Gallus gallus*, Aves; **Supp. Tab. S3.4**), and covers 12 major orders, while the fruit fly dataset consists of 17 species and 1 outgroup (*Anopheles gambiae*, Culicidae; **Supp. Tab. S5**), and covers 2 *Drosophila* subgenera (i.e., *Drosophila* and *Sophophora*). BUSCO statistics for complete single-copy genes were generally higher than those of bivalves, with a median of 98.3% for mammals and of 99.8% for fruit flies (**Supp. Tab. S3.4–S5**).

3.3.2 The Dmrt, Sox, and Fox complements in bivalves

Our annotation pipeline managed to successfully identify and annotate Dmrt, Sox, and Fox genes (DSFGs) in bivalves, as proved by the same analysis in mammals and fruit flies (see the paragraph **The Dmrt, Sox, and Fox complements and their amino acid divergence in the testing datasets** COMPLETARE COMPLETARE).

We retrieved four main orthology groups of *doublesex* and *mab-3* related transcription factor (Dmrt) genes in bivalves (**Fig. 3.2**; **Supp. Fig. S3.1**; **Supp. Tab. S6**), three corresponding to the groups present in the Bilateria common ancestor (*Dmrt-2*, *Dmrt-3*, and *Dmrt-4/5*; **Mawaribuchi et al., 2019**), and one additional group with no unambiguous ortholog among reference genes, and thus putatively specific to molluscs (named *doublesex and mab-3 related transcription factor 1-like* (*Dmrt-1L*), as per **Li et al., 2018; Evensen et al., 2022**). The majority of identified Dmrt genes are present in single-copy in each species, but *Dmrt-4/5*s show a group-specific expansion in Palaeoheterodonta and Heterodonta, while *Dmrt-1L* is completely absent from Heterodonta. The degree of missing data for Dmrt genes in bivalves is about 35%, with *Dmrt-2* having the highest (about 56%) and *Dmrt-4/5* the lowest (about 7%; **Supp. Tab. S7**). The coupling of ubiquitin conjugation to endoplasmic reticulum degradation (CUE)-like DM-associated (DMA) domain has been annotated in most of the *Dmrt-3* and *Dmrt-4/5* genes, while an additional *dsx* and *mab-3* (DM) domain has been annotated in *Dmrt-1L* genes in Mytilida and the gastropod *Pomacea canaliculata* (**Supp. Tab. S6**). Additionally, we retrieved six main orthology groups of *Sry*-related HMG-box (Sox) genes, none of which is restricted to molluscs or bivalves (**Fig. 3.2**; **Supp. Fig. S3.2**; **Supp. Tab. S6**). Five Sox groups (*Sox-B1/2*, *Sox-C*, *Sox-D*, *Sox-E*, and *Sox-F*) are those traditionally considered to be present in the Bilateria common ancestor (**Phochanukul and Russell, 2010**), while one has been identified outside mammals only recently (*Sox-H*, or *Sox-30*; **Han et al., 2010**). *Sox-B2* and *Sox-B1* have been grouped in the same clade, as in our phylogenetic reconstruction the former results in a paraphyletic group with the latter (**Supp. Fig. S3.2**), despite being traditionally recognised as a separate paralogy group in humans, fruit flies, and nematodes. The degree of missing data for Sox genes in bivalves is about 8%, with *Sox-H* having the highest (about 21%) and *Sox-B1/2* and *Sox-C* both having no missing genes (**Supp. Tab. S7**). The Sox N-terminal signature domain was annotated for *Sox-E* genes (**Supp. Tab. S6**). Concerning forkhead box (Fox) genes, we retrieved 27 main orthology groups (**Fig. 3.2**; **Supp. Fig. S3.3**; **Supp. Tab. S6**), two of which are specific to molluscs (*Fox-OG13/NA*,

Fox-OG16/NA). Additionally, other potential mollusc-specific Fox groups have been identified, but these have been excluded from the final orthology analysis as they are present in less than half of bivalve species (see **Materials and Methods** REFERENCE REFERENCE; **Supp. Tab. S6**). The two major Fox gene subgroups, Group I (monophyletic, specific to Metazoa; includes *Fox-A*, *Fox-B*, *Fox-C*, *Fox-D*, *Fox-E*, *Fox-F*, *Fox-G*, *Fox-H*, *Fox-L1*, *Fox-L2*, *Fox-Q2*) and Group II (paraphyletic, specific to Opisthokonta; includes *Fox-O*, *Fox-P*, *Fox-J2*, *Fox-J1*, *Fox-K*, *Fox-N2/3*, *Fox-N1/4*; **Larroux et al., 2008**), have been recovered, including the four Fox genes that were present in the Bilateria common ancestor (*Fox-C*, *Fox-F*, *Fox-L1*, and *Fox-Q1*; **Shimeld et al., 2010**). Two putative lineage-specific expansions have been recovered for *Fox-OG28/NA*, one regarding *Mytilus* spp. and one regarding the two Myida species (**Fig. 3.2**; **Supp. Fig. S3.3**). The degree of missing data for Fox genes in bivalves is about 22%, with *Fox-H* having the highest (about 42%) and *Fox-J1* having no missing genes (**Supp. Tab. S7**). The forkhead-associated (FHA) domain was annotated for *Fox-K* genes, the *Fox-P* coiled-coil signature domain was annotated for *Fox-P* genes, while both the forkhead N- and C-terminal signature domains were annotated for *Fox-A* genes (**Supp. Tab. S6**). Regarding bivalve species, the amount of missing data greatly differs between genomes and transcriptomes, with a mean of about 9% and about 45%, respectively. *Argopecten irradians concentricus*, *Mytilus unguiculatus* (formerly *coruscus*), and *Pecten maximus* have no missing data, while *Loripes orbiculatus* has the highest proportion (about 64%; **Supp. Tab. S7**).

3.3.3 Amino acid sequence divergence of Dmrt, Sox, and Fox genes in bivalves

In the reduced bivalve dataset, OrthoFinder collectively analysed >1.2G genes distributed in 34 species. 89.4% of these genes were placed in orthogroups, while 10.6% were not. The number of retrieved single-copy orthogroups (SCOs) is 5, which is drastically low but can be explained considering the mixed nature of the dataset, that is, it includes both genomes and transcriptomes with highly different BUSCO scores (**Supp. Tab. S3.1**). In order to be able to analyse a greater number of genes, we decomposed OrthoFinder orthogroups using DISCO and eventually obtained 11k SCOs with at least 50% of the species. By running the same pipeline on DSFGs, we included in the amino acid sequence divergence (AASD) analysis 32 SCOs (**Fig. 3.2**) out of 33 initial Possvm-identified groups (*Fox-H* didn't meet the species occupancy threshold; **Fig. 3.3**).

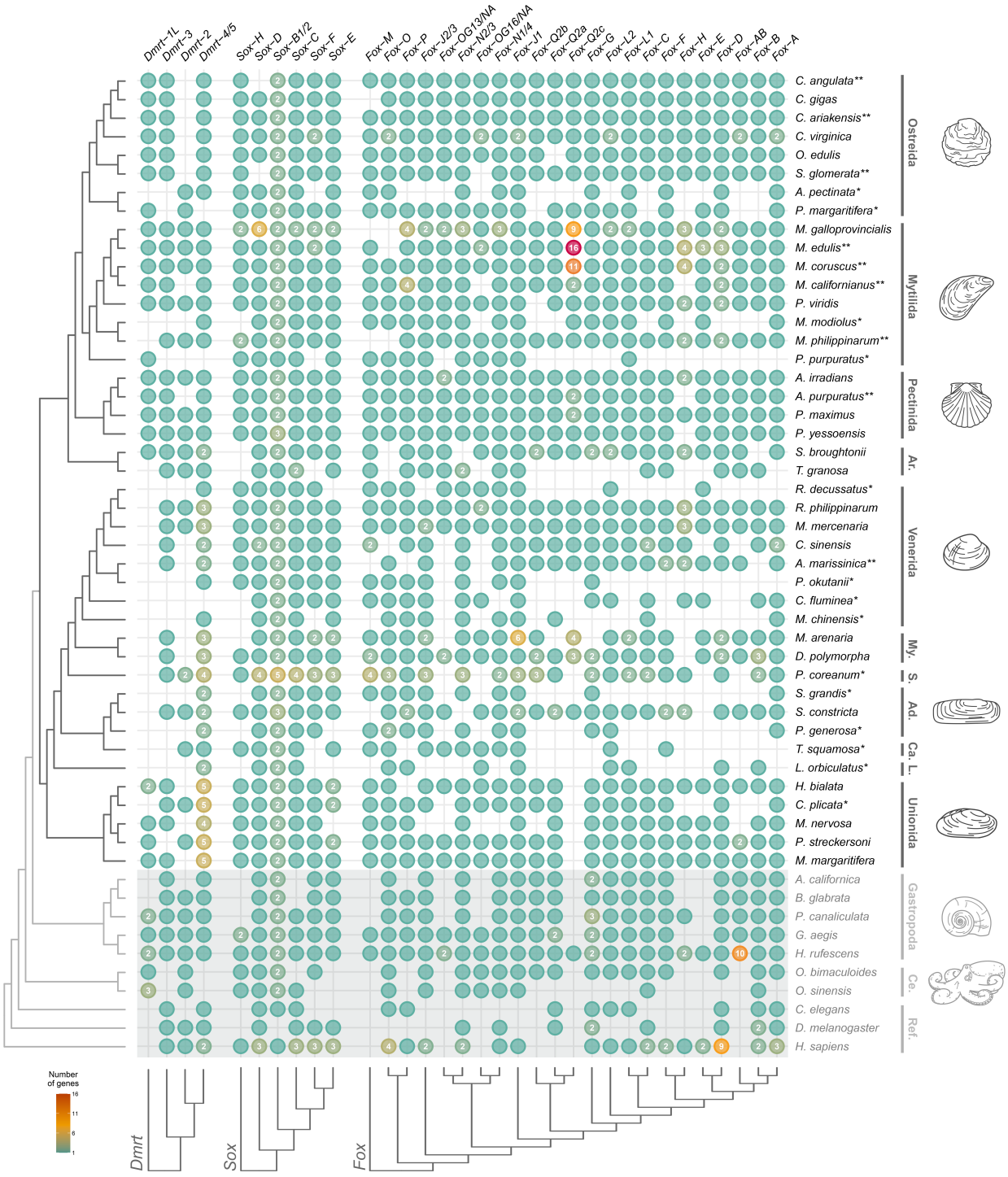


Figure 3.2. DSFG complement in bivalves and their outgroups. Presence/absence of genes in various species are indicated by filled circles. Numbers inside each circle specify genes with 2 or more copies. The shaded area highlights non-bivalve species, belonging either to other molluscs or to the references. The phylogenetic tree of analyzed species, as inferred from literature, is shown on the left, while major taxonomic groups are reported on the right. Species represented by transcriptomic data are marked with an asterisk ('*'), and species not present in the reduced bivalve dataset are marked with two asterisks ('**'; see main text and **Fig. 3.1**); note that the two categories do nor overlap. DSFG trees are shown on the bottom (full trees can be found in **Supp. Fig. S3.1**–**S3.3**). Full species names, along with all assembly and taxonomic information, can be found in **Supp. Tab. S3.1**. Ad.: Adapedonta; Ar.: Arcida; Ca.: Cardiida; Ce.: Cephalopoda; L.: Lucinida; My.: Myida; Ref.: reference genes; S.: Sphaeriida.

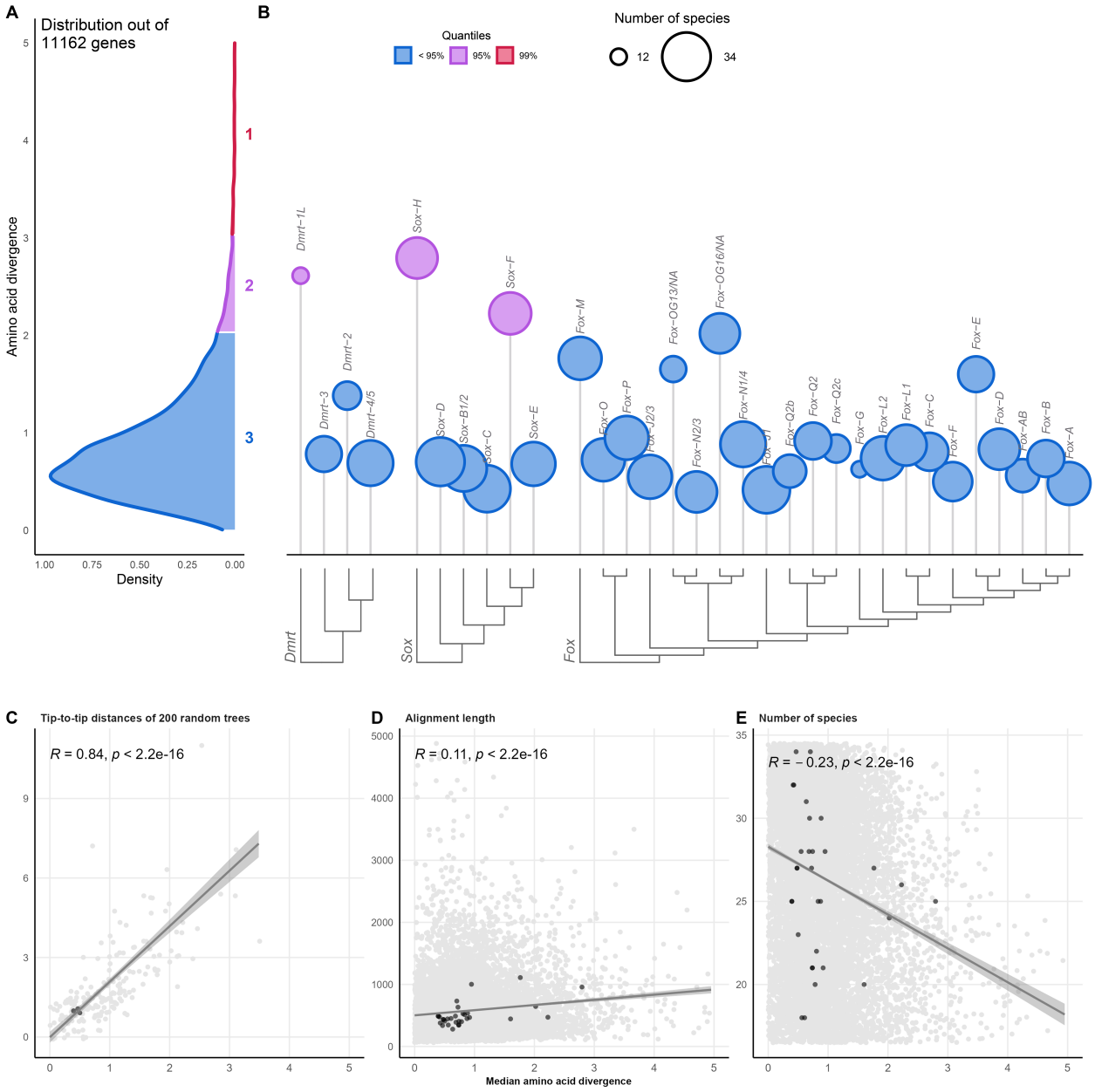


Figure 3.3. Distribution of AASD of single-copy orthogroups in bivalves (A), including DSFGs (B), and their correlations with tip-to-tip distances (C), alignment lengths (D), and number of species (E). The distribution of AASD has been computed on the median values of pairwise distances of >11k SCOs from the reduced bivalve dataset (see main text and **Fig. 3.1**). Genes have been divided according to their median AASD value into three different groups, which are indicated by different colors and increasing numbers (Groups 1, 2, and 3). Circle heights of DSFGs show the median value of their AASD, while the size indicates the number of represented species. DSFG trees are shown on the bottom (full trees can be found in **Supp. Fig. S3.1–S3.3**). Darker points in C–E indicate DSFG SCOs. The correlation between the amino acid distance and the tip-to-tip distance has been computed on 200 randomly-selected orthogroups.

From the distribution of median AASD, 112 genes were assigned to Group 1 (1% upper quantile), 447 to Group 2 (5% upper quantile), and 10.603 to Group 3. Most of the DSFGs (29/32) fell in Group 3 (**Fig. 3.3**), which means they have a median AASD comparable to the vast majority of other genes in bivalves (median level of the genomes). Just *Dmrt-1L*, *Sox-H*, and *Sox-F* showed higher divergences, and have been accordingly placed in Group 2. Overall, pairwise AASD proved to be a good approximation of the tip-to-tip distances ($R = 0.84, p < 2.2\text{e}{-16}$, calculated on 200 randomly-selected trees; **Fig. 3.3C**), while it showed no influence from the alignment length ($R = 0.11$) or the number of represented species ($R = -0.23$; **Fig. 3.3D–E**). Genes from Group 1 and Group 2 are strongly involved in cellular regulatory processes (such as those related to the metabolism of nucleic acids, proteins, and other macromolecules), but also in development and response to external stimuli, as shown by the GO-term enrichment analysis (**Tab. 3.1**; **Supp. Tab. S10**).

Table 3.1. Enriched GO terms for Group 1 and Group 2 genes of bivalves, mammals, and *Drosophila*. The extended version of the table, which includes also the expected number of annotated genes per GO term and all the other enriched GO terms, can be accessed in **Supp. Tab. S10**.

Dataset	GO.ID	Term	Annotated genes	Significant genes	Corrected p-value
	GO:0060255	regulation of macromolecule metabolic process	737	59	0.04525
	GO:0080090	regulation of primary metabolic process	673	53	0.01818
	GO:0019219	regulation of nucleobase-containing compound metabolic process	541	41	0.02388
	GO:0006351	DNA-templated transcription	571	39	0.03767
	GO:0032774	RNA biosynthetic process	579	39	0.04490
	GO:0051252	regulation of RNA metabolic process	517	37	0.02719
	GO:0006355	regulation of DNA-templated transcription	490	35	0.03751
	GO:2001141	regulation of RNA biosynthetic process	491	35	0.03844
	GO:0006950	response to stress	370	33	0.01949
	GO:0032502	developmental process	261	27	0.04445
Bivalvia	GO:0006468	protein phosphorylation	345	23	0.02483
	GO:0031325	positive regulation of cellular metabolic process	125	17	0.00801
	GO:0010604	positive regulation of macromolecule metabolic process	151	17	0.04047
	GO:0051172	negative regulation of nitrogen compound metabolic process	117	16	0.00814
	GO:0051173	positive regulation of nitrogen compound metabolic process	137	15	0.02454
	GO:0006310	DNA recombination	66	14	0.00087
	GO:0048513	animal organ development	83	12	0.04088
	GO:0010629	negative regulation of gene expression	78	11	0.00048
	GO:0023051	regulation of signaling	133	11	0.02872
	GO:0045934	negative regulation of nucleobase-containing compound metabolic process	64	11	0.03637
	GO:0009605	response to external stimulus	90	11	0.04544

Tab. 3.1 continued from previous page

Dataset	GO.ID	Term	Annotated genes	Significant genes	Corrected p-value
Bivalvia	GO:0044419	biological process involved in interspecies interaction between organisms	63	11	0.04761
Mammalia	GO:0006955	immune response	1297	145	0.00061
	GO:0098542	defense response to other organism	853	112	0.02066
	GO:0045087	innate immune response	647	82	8.5e-10
	GO:0001817	regulation of cytokine production	630	51	0.04660
	GO:0042742	defense response to bacterium	233	45	1.7e-07
	GO:0006954	inflammatory response	642	45	0.01735
	GO:0019221	cytokine-mediated signaling pathway	382	44	3.9e-07
	GO:0002250	adaptive immune response	342	44	1.3e-05
	GO:0001819	positive regulation of cytokine production	402	41	0.02723
	GO:0002697	regulation of immune effector process	308	37	0.04426
	GO:0042110	T cell activation	432	35	0.02564
	GO:0051607	defense response to virus	257	34	1.9e-07
	GO:0048232	male gamete generation	491	32	0.02255
	GO:0007283	spermatogenesis	478	31	0.02801
	GO:0070661	leukocyte proliferation	273	29	0.01285
	GO:0002449	lymphocyte mediated immunity	221	29	0.04833
	GO:0070663	regulation of leukocyte proliferation	212	25	0.01870
	GO:0050727	regulation of inflammatory response	300	24	0.00235
	GO:0031349	positive regulation of defense response	240	24	0.01239
	GO:0002768	immune response-regulating cell surface receptor signaling pathway	177	22	0.00336
	GO:0050829	defense response to Gram-negative bacterium	66	17	1.7e-10
	GO:0071222	cellular response to lipopolysaccharide	164	17	0.00012

Tab. 3.1 continued from previous page

Dataset	GO.ID	Term	Annotated genes	Significant genes	Corrected p-value
	GO:0010466	negative regulation of peptidase activity	163	16	0.00036
	GO:0002429	immune response-activating cell surface receptor signaling pathway	164	16	0.00243
	GO:1903555	regulation of tumor necrosis factor superfamily cytokine production	137	16	0.01244
	GO:0071706	tumor necrosis factor superfamily cytokine production	137	16	0.01244
	GO:0070665	positive regulation of leukocyte proliferation	132	16	0.02765
	GO:0045089	positive regulation of innate immune response	113	16	0.03224
Mammalia	GO:0071356	cellular response to tumor necrosis factor	175	15	0.00219
	GO:0002695	negative regulation of leukocyte activation	148	15	0.01151
	GO:0002456	T cell mediated immunity	82	15	0.01605
	GO:0002705	positive regulation of leukocyte mediated immunity	113	15	0.01837
	GO:0032680	regulation of tumor necrosis factor production	133	15	0.03262
	GO:0032640	tumor necrosis factor production	133	15	0.03262
	GO:0050866	negative regulation of cell activation	165	15	0.04048
	GO:0000819	sister chromatid segregation	140	11	0.02927
	GO:0070192	chromosome organization involved in meiotic cell cycle	54	9	0.00849
	GO:0007131	reciprocal meiotic recombination	37	7	0.00066
	GO:0007143	female meiotic nuclear division	54	6	0.02270
	GO:0035967	cellular response to topologically incorrect protein	44	5	0.03334
Drosophila	GO:0035966	response to topologically incorrect protein	47	5	0.04266
	GO:0007141	male meiosis I	13	4	0.00150
	GO:0140543	positive regulation of piRNA transcription	3	3	6.9e-05
	GO:0010526	retrotransposon silencing	8	3	0.00331
	GO:0007130	synaptonemal complex assembly	10	3	0.00666

Tab. 3.1 continued from previous page

Dataset	GO.ID	Term	Annotated genes	Significant genes	Corrected p-value
	GO:0030719	P granule organization	11	3	0.00888
	GO:0071218	cellular response to misfolded protein	12	3	0.01149
Drosophila	GO:0051788	response to misfolded protein	12	3	0.01149
	GO:0007135	meiosis II	15	3	0.02169
	GO:0034508	centromere complex assembly	19	3	0.04094

3.3.4 Dmrt, Sox, and Fox genes, and amino acid sequence divergence in the test datasets

The DSFG datasets retrieved in mammals and fruit flies are far more complete than those in bivalves, and most of the already-recognised orthology groups have been identified.

In mammals, we retrieved 7 Dmrt orthology groups with about 3.1% of missing data, 20 Sox orthology groups with about 8.1% of missing data, and 42 Fox orthology groups with about 4.6% of missing data (**Supp. Fig. S3.4A, S3.5–S3.7**; **Supp. Tab. S8**). Of these, just *Sox-5* was not included in the subsequent AASD analysis, as it did not meet the 50%-species occupancy threshold. OrthoFinder analysed about 650M genes, and the number of SCOs used in the AASD analysis (thus resulting from the DISCO-based orthogroup decomposition pipeline) is >16k (**Fig. 3.4A**). From the distribution of median AASD, 163 genes were assigned to Group 1, 649 to Group 2, and 15.355 to Group 3. Most of the DSFGs (66/68) fell in Group 3 (**Fig. 3.4B**), while *Sex-determining region of chromosome Y* (*Sry*) and *Fox-D4* showed higher divergences, and have been accordingly placed in Group 1 and 2, respectively. Genes from Group 1 and Group 2 show a strong enrichment in immune-related functions (such as innate and adaptive immune response, defence response to bacteria and viruses, lymphocyte methabolism, etc.), but also in reproductive processes (such as spermatogenesis; **Tab. 3.1**; **Supp. Tab. S10**).

Concerning *Drosophila*, we retrieved 4 Dmrt orthology groups with about 1.7% of missing data, 7 Sox orthology groups with about 3.9% of missing data, and 17 Fox genes with about 8.3% of missing data (**Supp. Fig. S3.4B, S3.8–S3.10**; **Supp. Tab. S9**). OrthoFinder analysed about 240M, and the distribution of median AASD was built after >12k SCOs (**Fig. 3.4C**). 126 genes were assigned to Group 1, 501 to Group 2, and 11.880 to Group 3. All of the DSFGs have been used in the AASD analysis, but none of them have been placed in Group 1 or 2, that is, all the DSFGs in *Drosophila* have an AASD comparable to the median level of the genome (**Fig. 3.4D**). Genes of Group 1 and Group 2 show a GO-term enrichment in meiotic processes, such as chromosome/chromatid organisation, and retrotransposon silencing (**Tab. 3.1**; **Supp. Tab. S10**).

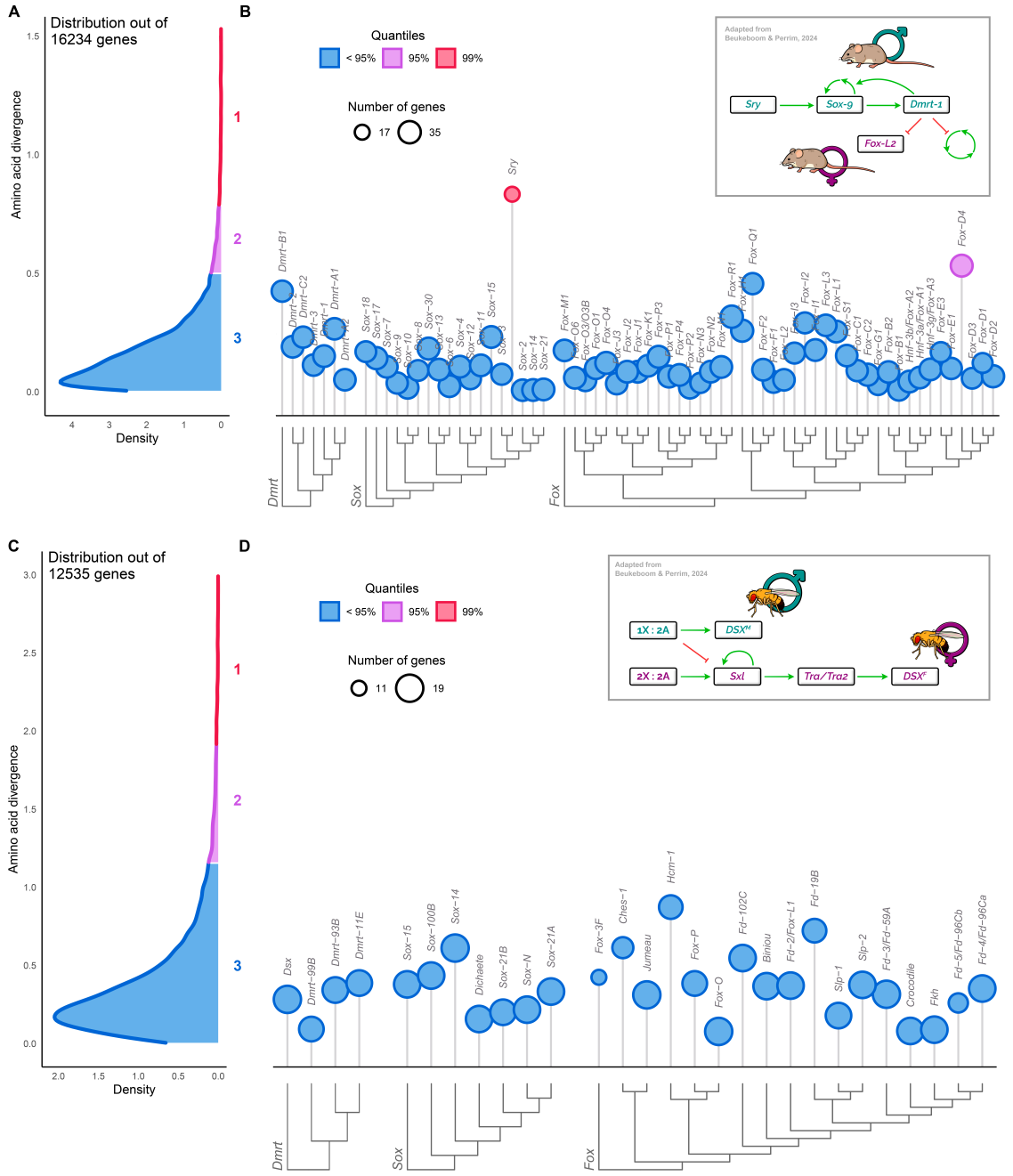


Figure 3.4. Distribution of amino acid divergence (AASD) of single-copy orthogroups in Mammalia (A) and *Drosophila* (C), including Dmrt, Sox, and Fox genes (DSFGs; B-D). The distributions of AASD in mammals and fruit flies have been computed on the median values of pairwise distances of over 16k and 12k SCOs, respectively. Genes have been divided according to their median AASD value into three different groups, which are indicated by different colors and increasing numbers (Groups 1, 2, and 3). Circle heights of DSFGs show the median value of their AASD, while the size indicates the number of represented species. DSFG trees are shown on the bottom (full trees can be found in **Supp. Fig. S3.5–S3.7** for mammals and in **Supp. Fig. S3.8–S3.10** for fruit flies). Insets: scheme of the sex-determination molecular pathways in *Mus musculus* and in *Drosophila melanogaster*, with shown the main genes involved (adapted from **Beukeboom and Perrin, 2014**). Green arrows indicate transcription activations, red arrows indicate transcription suppressions. X: sex chromosomes; A: autosomal chromosomes; $DSX^{M/F}$: *DSX* splicing variants present in males or females, respectively.

3.4 Discussion

3.4.1 A new manually-curated and phylogenetic-based reference dataset of Dmrt, Sox, and Fox genes in bivalves

The annotation and characterization process of a gene family in a certain clade of organisms may harbour many overlooked challenges (**Vizueta Moraga et al., 2020**). For example, the presence of highly-conserved catalytic domains may hamper the correct identification of the components of a gene family, as it is the case for *Hox* and *ParaHox* genes and their homeobox motif (**Baldwin-Brown et al., 2018; Nicolini et al., 2023**). Conversely, the components of dynamic gene families characterised by abrupt and sequential duplication events may be difficult to sort into separate groups. As a matter of fact, varying levels of sequence heterogeneity and of gene copy numbers makes the inference of orthologous groups hard, as for certain P450 clans (**Dermauw et al., 2020**). Regardless of the causes, having a solid and wide phylogenetic context in which to study gene duplications and losses, and orthology relationships, is crucial to overcome these difficulties. In the same way, manual curation and visual inspection of multiple sequence alignments, phylogenetic trees, and gene structures (in terms of domain annotation, start and stop codons, and other feature representations) is helpful, despite being time-demanding and possibly low reproducible. In this study, we characterised the full complement of DSFGs in the vast clade of bivalves, by leveraging sequence domain annotation, phylogenetics, and manual curation of the dataset. Our aim was to obtain the most reliable gene complements as possible, combined with a vast taxonomic dataset, a solid phylogenetic inference, an openly-available dataset of gene sequences, and a reproducible pipeline for the annotation of gene identity. By doing so, we want to provide a reliable resource for future studies of DSFGs, either focused on bivalves or generally in Metazoa.

Our approach allowed us to identify some cases of incorrect gene identification and relative nomenclature in bivalves, which have arisen because of erroneous or ambiguous annotations in previous works, as a result of limited datasets or analyses. For example, concerning the Dmrt gene family, we identified orthologs of the vertebrate *Dmrt-2*, *Dmrt-3*, and *Dmrt-4/5* (*A1/A2*; **Fig. 3.2**; **Supp. Fig. S1**; **Supp. Tab. S6**), which are also expected to have been present in the Bilateria common ancestor (**Mawaribuchi et al., 2019**). **Wang et al., 2023** found that *Dmrt-4/5* is duplicated in *Mercenaria mercenaria* and *Cyclina sinensis* (Venerida), and in *Dreissena polymorpha* (Myida), and we confirm this result by tracing back the duplication event

to the split between Palaeoheterdonta (here represented by Unionida) and Heterodonta (here represented by Venerida, Myida, Sphaeriida, Adapedonta, Cardiida, and Lucinida; **Fig. 3.2**). Furthermore, we confirm *Dmrt-1L* to be present in many bivalve species (mainly belonging to the Ostreida, Pectinida, Mytilida, and Unionida orders; **Fig. 3.2**), as well as in gastropods and *Octopus*. Though, our phylogenetic analysis did not retrieve any unambiguous orthology relationship among *Dmrt-1L* and either vertebrate *Dmrt-1* or *Drosophila dsx* genes, as instead it was proposed in previous works (**Li et al., 2018; Evensen et al., 2022**). As a matter of fact, the amino acid sequence of the *Dmrt-1L* DM domain does not recall that of any other Dmrt gene. Furthermore, it must be considered that various phylogenetic analyses have recovered both *Dmrt-1* and *dsx* genes to be restricted to vertebrates and arthropods, respectively (**Wexler et al., 2014; Mawaribuchi et al., 2019; Panara et al., 2019**), that is, they do not have any direct ortholog outside their relative clades. Thus, if *Dmrt-1L*, *dsx*, and *Dmrt-1* are true orthologs, their origin would need to be placed at least in the Bilateria common ancestor, which seems however to not be the case. All considered, we thus confirm that *Dmrt-1L* is not homologous to *Dmrt-1* and *Dsx* and is rather a mollusc-specific gene (**Evensen et al., 2022**). The monophyly of the group, which is not supported by the phylogenetic tree inferred with Dmrt genes from also the reference species (**Supp. Fig. S1**), is instead recovered when analysing just genes from mollusc species (Supp Fig. S11). To this regard, we speculate that in our analysis, the difficulty in obtaining the monophyly of *Dmrt-1L* genes may have arisen primarily because of the many *C. elegans*-restricted genes (**Supp. Tab. S3.3**), which are placed among the other bivalve genes (**Supp. Fig. S1**), but also because of the high AASD of *Dmrt-1L* genes (see High amino acid sequence divergence and identification of sex-determining genes LINK LINK LINK), which hampers a straight-forward phylogenetic reconstruction. Our broad-context analysis also suggests that (i) the scallop-specific cluster of Dmrt genes retrieved by **Wang et al., 2023** rather belongs to the *Dmrt-1L* group, and (ii) the classification of Dmrt genes of bivalves provided by **Zeng et al., 2024** needs to be revised as proposed in this work (for example, *Dmrt-1* genes are *Dmrt-4/5*; *Dmrt-2* genes are *Dmrt-3*; *Dmrt-3* genes are *Dmrt-1L*; hence, Crassostrea species do not have *Dmrt-2* genes).

For what concerns the Sox gene family, bivalves (or molluscs) do not show any major clade-restricted gene, as only the five Bilateria-specific Sox groups (*Sox-B1/2*, *Sox-C*, *Sox-D*, *Sox-E*, and *Sox-F*) and *Sox-H* have been mainly identified (**Fig. 3.2**; **Supp. Fig. S2**; **Supp. Tab. S6**), in accordance with previous findings (**Evensen et al., 2022; Wang and Nie, 2024;**

Yu, Zhang, et al., 2017). *Sox-B1/2* is clearly made up of two subgroups (i.e., *Sox-B1* and *Sox-B2*), as expected, but their respective identity could not be unambiguously established, as *Sox-B1/2* genes of reference species do not form separate clusters (**Supp. Fig. S2**). Even when inferring the phylogenetic tree only of components of the *Sox-B1/2* group from molluscs and reference species, the identity can not be established properly (**Supp. Fig. S12**).

Compared to Dmrt and Sox genes, the Fox gene family appears as the most dynamic in terms of gene presence/absence, as already shown by other works (**Wu et al., 2020; Schomburg et al., 2022; Seudre et al., 2022**). Our phylogenetic analysis successfully recovered Group I and Group II of Fox genes (**Larroux et al., 2008**), which include the four Fox genes that were present in the Bilateria common ancestor (*Fox-C*, *Fox-F*, *Fox-L1*, and *Fox-Q1*; **Fig. 3.2**; **Supp. Fig. S3**; **Supp. Tab. S6**; **Shimeld et al., 2010**). To our knowledge, this is the first broad-taxonomic identification and classification of Fox genes in bivalves, as up to now they have been systematically characterised only in *Crassostrea gigas* (**Yang et al., 2014**), *Patinopecten yessoensis* (**Wu et al., 2020**), and *Ruditapes philippinarum* (**Liu et al., 2024**). Firstly, our analysis confirms the absence in molluscs of *Fox-I*, *Fox-Q1*, *Fox-R*, *Fox-S* (**Supp. Fig. S3**), which are in fact thought to have emerged with the diversification of deuterostomes or vertebrates (**Yang et al., 2014; Wu et al., 2020; Schomburg et al., 2022; Seudre et al., 2022**). Furthermore, we found many Fox groups that appeared as mollusc-specific and/or still-unnamed at a first analysis. However, a much more in-depth investigation revealed a different scenario. *Fox-OG2/NA* appears close to the human *Fox-M* gene in the phylogenetic tree, but they do not form a monophyletic group (**Supp. Fig. S3**). However, by comparing *Fox-OG2/NA* sequences and the phylogenetic tree with those analysed by **Yang et al., 2014**, **Wu et al., 2020**, **Schomburg et al., 2022**, **Seudre et al., 2022**, it appears clear that this group of Fox genes is indeed *Fox-M*. However, our analysis has failed to retrieve a monophyletic relationship among bivalve and human *Fox-M* genes, even when inferring a tree with just *Fox-J2*, *Fox-M*, *Fox-O*, and *Fox-P* complements (**Supp. Fig. S13**), which belong to the same Fox group. Regarding the *Fox-OG39/NA* group, it does not have any homolog in reference species (**Supp. Fig. S3**) but is found to belong to the *Fox-AB* group by sequence comparison with previous works (**Yang et al., 2014; Wu et al., 2020; Seudre et al., 2022**). *Fox-AB* was formerly described only in the sea urchin *Strongylocentrotus purpuratus* and the lancelet *Branchiostoma floridae* (**Tu et al., 2006; Yu et al., 2008**), but was later identified also in several Spiralia lineages, including molluscs (e.g., **Yang et al., 2014; Wu et al., 2020;**

Seudre et al., 2022). A similar situation concerns *Fox-OG15/NA* and *Fox-OG28/NA*, which again could not be named based on orthology relationships with the reference species genes (**Supp. Fig. S3**), but actually represent two lineage-specific expansions of the *Fox-Q2* group (named *Fox-Q2b* and *Fox-Q2c*), as already appointed in previous studies (**Yang et al., 2014; Wu et al., 2020**). This observation fits within the wider context of the *Fox-Q2* group expansion in Bilateria and, particularly, in Spiralia, that led to remarkable differences in their gene copy numbers across various clades (**Seudre et al., 2022**). Two additional Fox genes have been previously identified in bivalves, and were named *Sox-Y* and *Sox-Z* (**Yang et al., 2014; Wu et al., 2020**). In our analysis, these Fox groups were identified as *Fox-OG13/NA* and *Fox-OG16/NA*, thanks to sequence comparison of Fox genes from *C. gigas* and *P. yessoensis*. On one hand, *Fox-Y* was firstly identified in *S. purpuratus* (**Tu et al., 2006**) and only recently in a few bivalve species (**Yang et al., 2014; Wu et al., 2020**). However, when analysing bivalve and *S. purpuratus* Fox genes, we failed in retrieving such a clear orthology relationship, as *S. purpuratus Fox-Y* does not fall within the phylogenetic range of bivalve *Fox-OG13/NA*, which contains the supposed *Fox-Y* orthologs (**Supp. Fig. S14**). Also, the forkhead domains of *Fox-OG13/NA* genes were annotated as 'forkhead domain P' (**Supp. Fig. S6**). On the other hand, *Fox-Z* was firstly identified in bivalves and in several other protostomes, thanks to a phylogenetic work including the brachiopod *Lingula unguis*, the annelid *Capitella teleta*, the scorpion *Centruroides sculpturatus*, and the centipede *Strigamia maritima* (**Wu et al., 2020**). However, later works have not recovered this result, even when analysing annelids (**Seudre et al., 2022**) and panarthropods (**Schomburg et al., 2022**) in a more focused effort. In this case, the forkhead domains were annotated as either a generic 'forkhead domain' or a 'forkhead domain Q2' (**Supp. Fig. S6**). All considered, we argue that bivalves possess two additional Fox groups (here *Fox-OG13/NA* and *Fox-OG16/NA*; **Fig. 3.2**; **Supp. Fig. S3**; **Supp. Tab. S6**) which are shared with other mollusc species, as revealed also by other authors. However, given the discordant results of the phylogenetic hypothesis and domain annotation, we think that a more thorough investigation on their orthology relationships with Fox genes from other Metazoa is needed, and thus we chose to not employ their former names *Fox-Y* and *Fox-Z*.

Besides the DSFG groups discussed so far, it must be also considered that many orphan genes have been identified (**Supp. Fig. S1–S3**; **Supp. Tab. S6**). For example, **Wu et al., 2020** identified a duplication event of *Fox-H* genes in *C. gigas*, which has been recovered also in our analysis for the entire Ostreida clade (*Fox-OG36/NA*; **Supp. Fig. S3**). Similarly, a

gene orthology group putatively specific to Pteriomorphia has been identified among Sox genes (*Sox-OG1/NA*). Of course, these genes deserve as much attention as the others, as they may constitute true group-specific expansions and may play fundamental roles in some biological processes. However, they have not been discussed here or included in **Fig. 3.2** for clarity purposes, but they are freely available in supplementary materials.

Overall, our analysis clearly shows the importance of adopting a wide-angle approach when characterising the members of a gene family, especially for large ones such as the Fox genes (**Schomburg et al., 2022**). As a matter of fact, the presence of duplication events and orphan genes needs to be addressed with a broad taxonomic dataset, in order to account for possible mis-annotations, gene phylogenetic mis-placements, and sequence heterogeneity. Additionally, many reference species need to be included for the gene identification process, in order to consider distantly-related genes and obtain a solid annotation. Our gene annotation pipeline also resulted to be very solid, even with non-model organisms and sub-optimal genomic and transcriptomic resources as they are those of bivalves. As a matter of fact, by running the same pipeline on two additional datasets composed of mammal and fruit fly genomes, we were able to obtain high-quality orthology groups in accordance with previous knowledge on the clades (**Supp. Fig. S4–S10**; **Supp. Tab. S8–S9**), with little or no manual curation. Furthermore, this is again the first broad analysis of DSFGs in both mammals and fruit flies, as so far attention has been dedicated only to single well-studied organisms (e.g., **Jackson et al., 2010**).

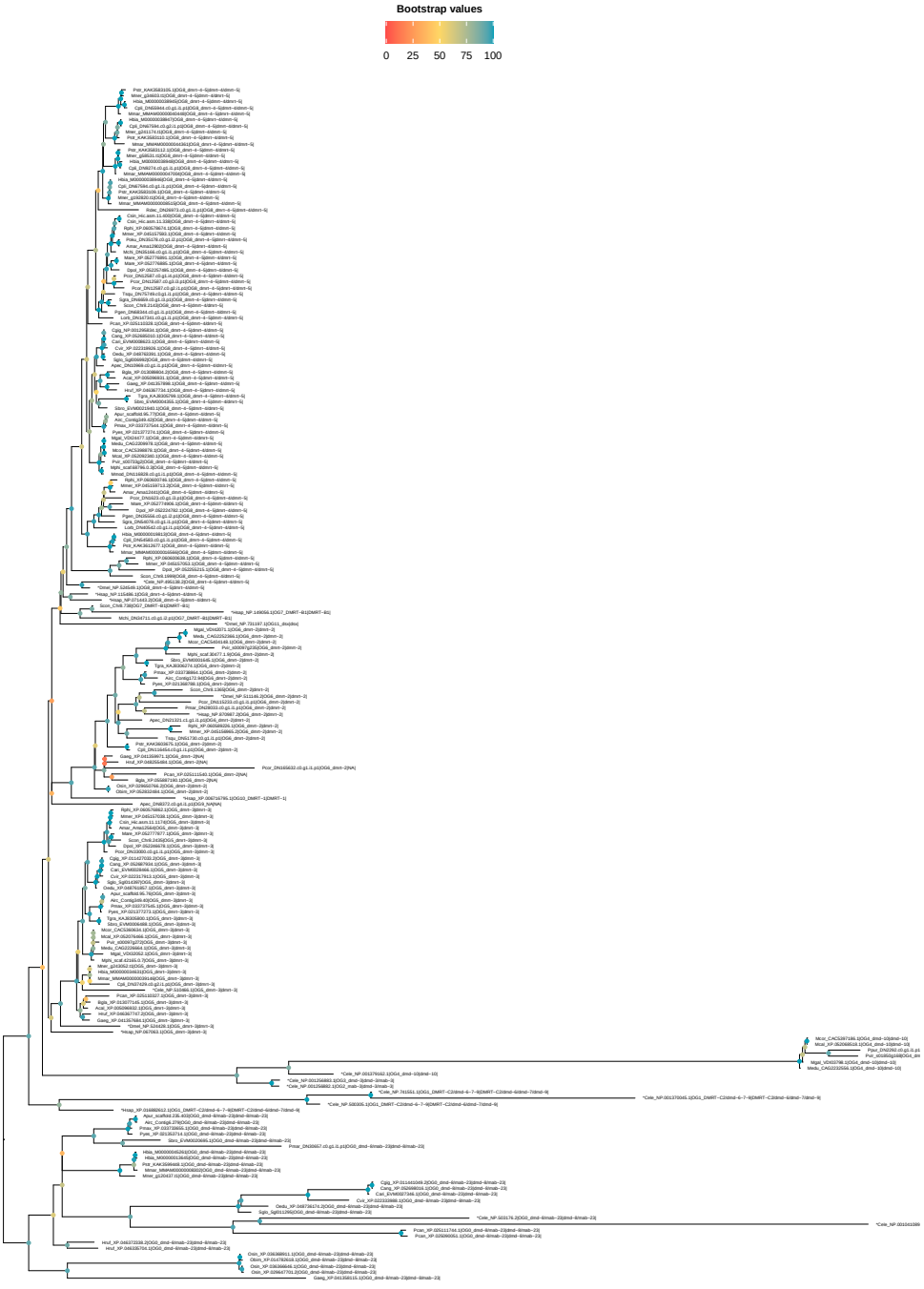
3.5 Conclusions.

In preparation.

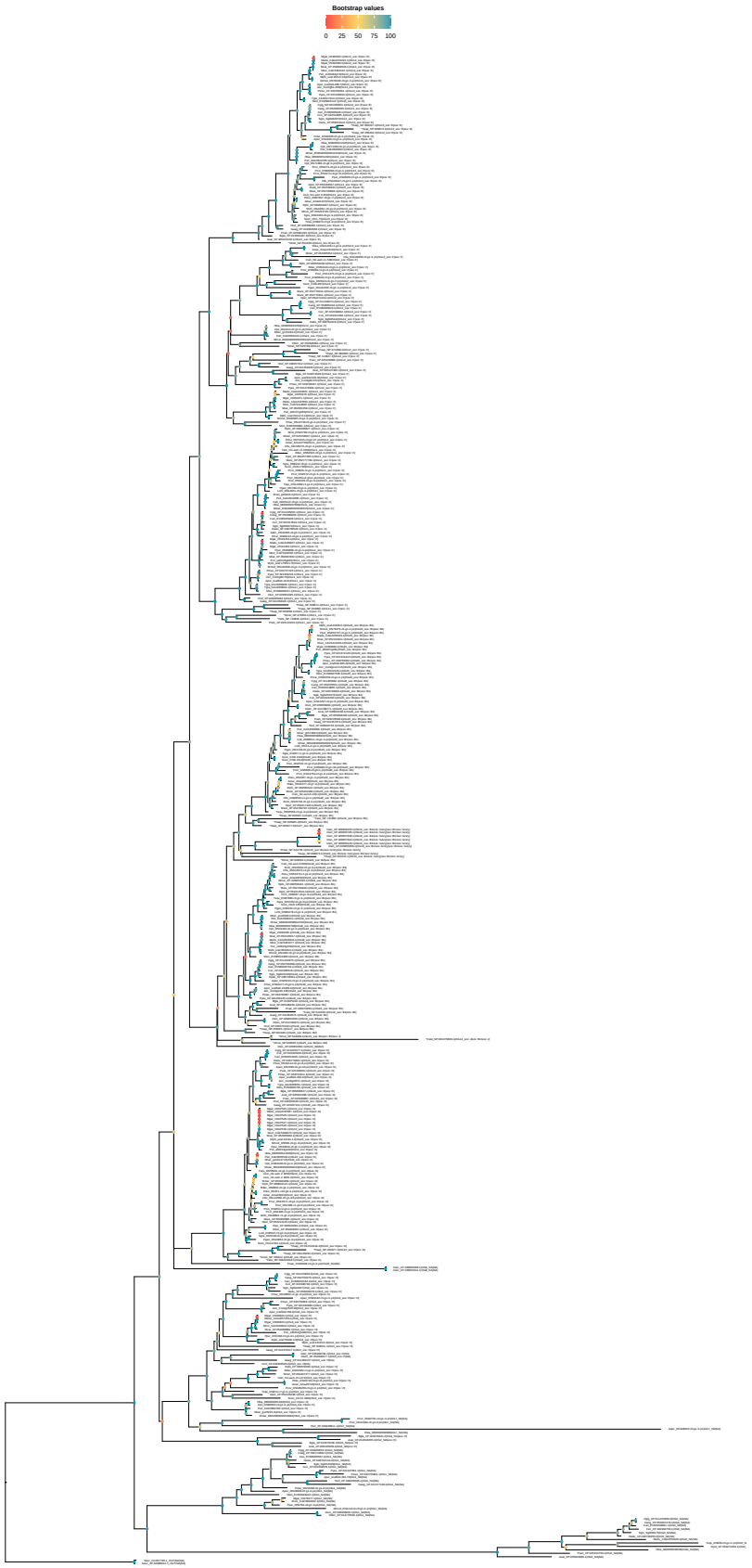
3.6 Supplementary Materials

All the supplementary materials will be available at my GitHub personal page.

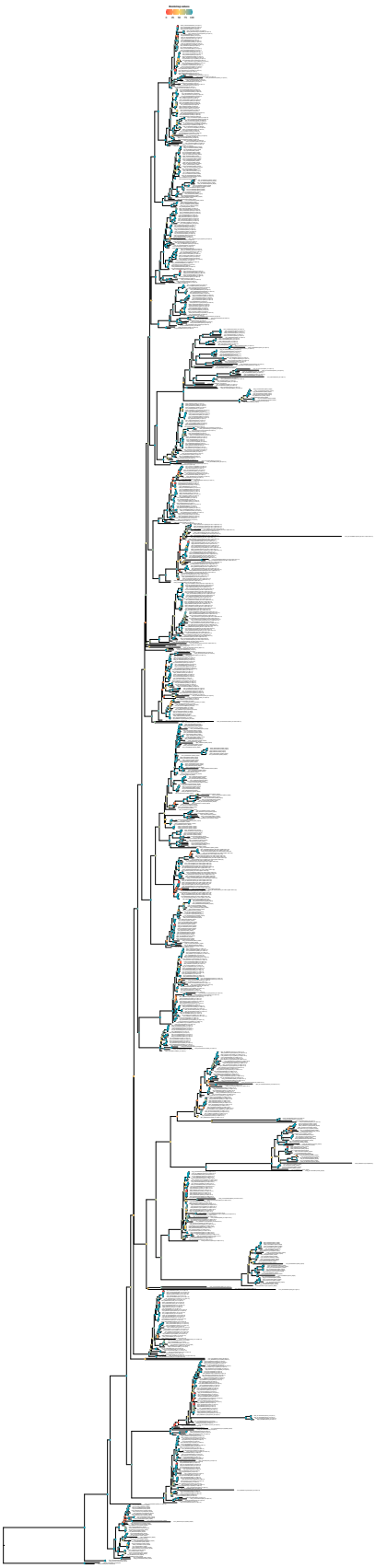
3.6.1 Supplementary Figures



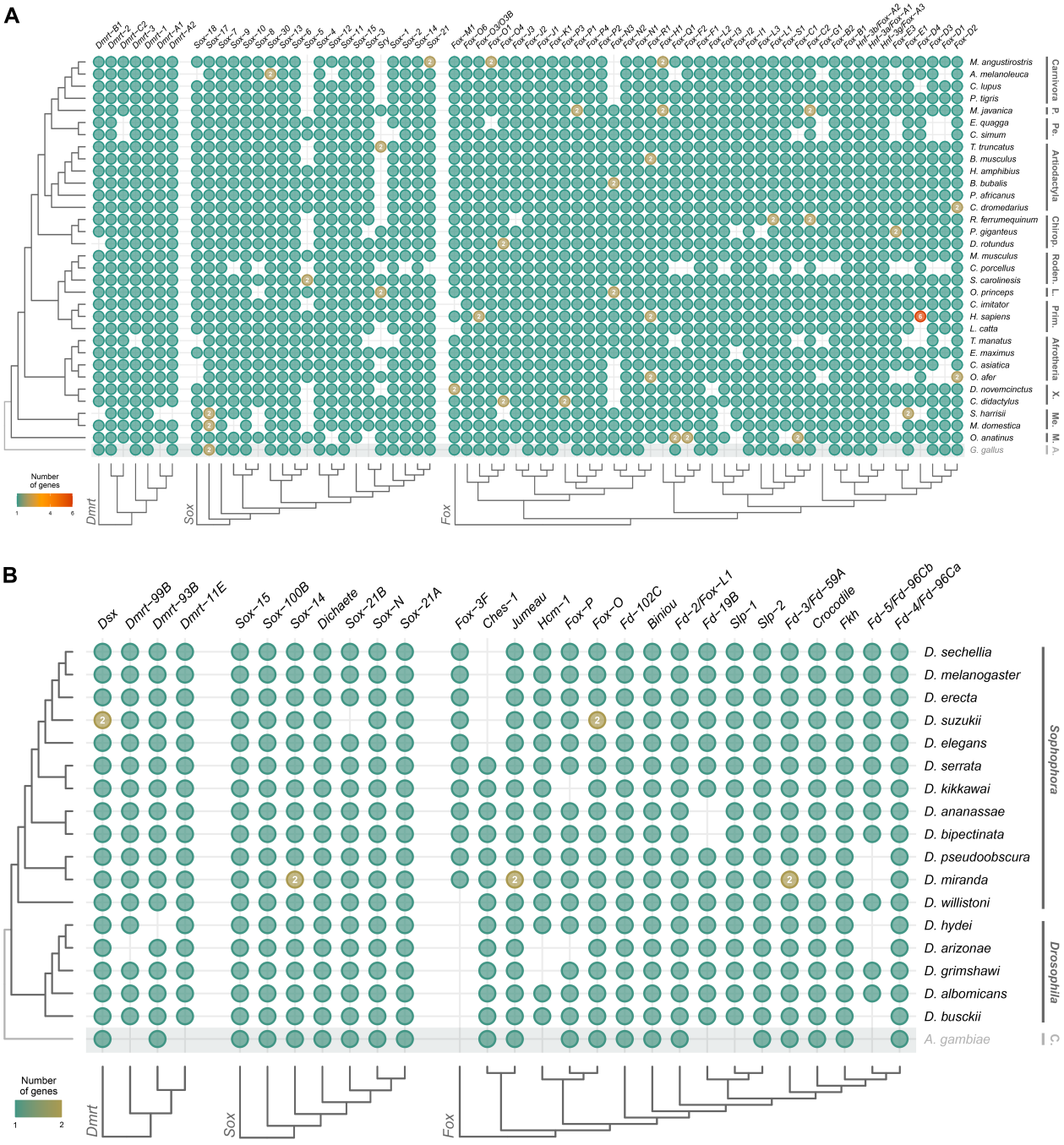
Supplementary Figure S3.1. ML phylogenetic tree of the Dmrt gene family in molluscs, including the possvm orthology inference. Reference genes from *Homo sapiens*, *Caenorhabditis elegans*, and *Drosophila melanogaster* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.1**. The tree has been midpoint rooted. Bootstrap values are shown for each node.



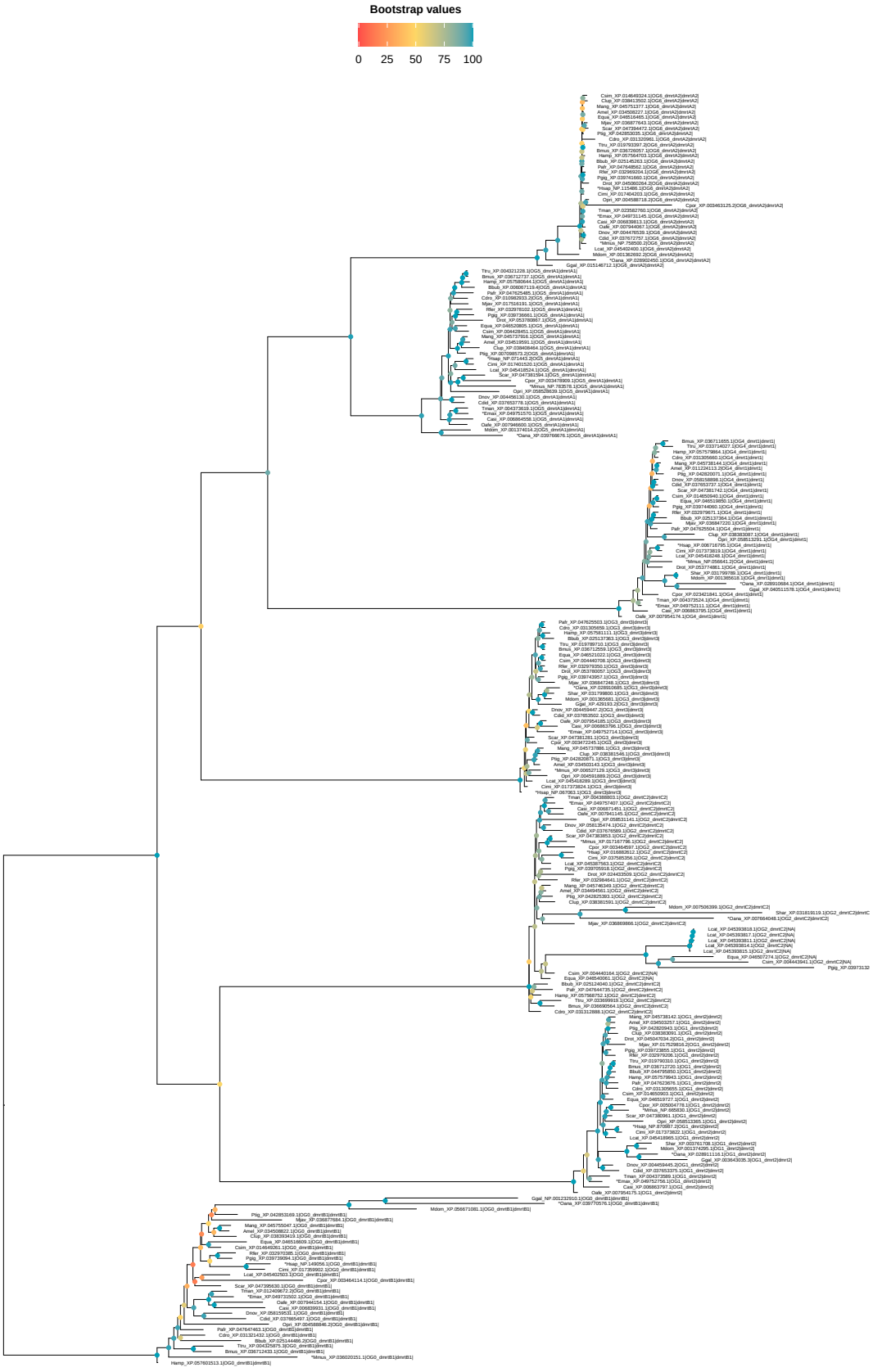
Supplementary Figure S3.2. ML phylogenetic tree of the Sox gene family in molluscs, including the possvm orthology inference. Reference genes from *H. sapiens*, *C. elegans*, and *D. melanogaster* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.1**. Bootstrap values are shown for each node.



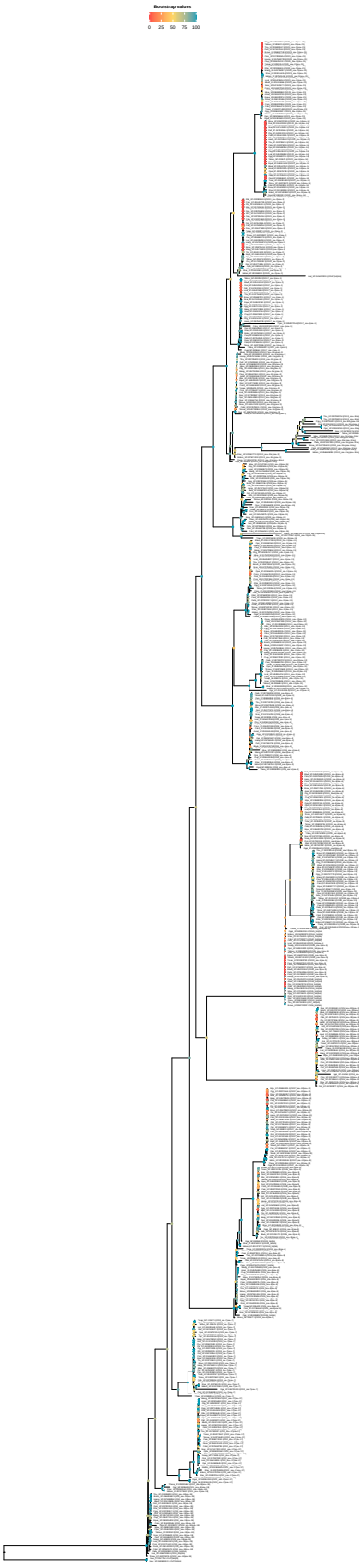
Supplementary Figure S3.3. ML phylogenetic tree of the Fox gene family in molluscs, including the possvm orthology inference. Reference genes from *H. sapiens*, *C. elegans*, and *D. melanogaster* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.1**. Bootstrap values are shown for each node.



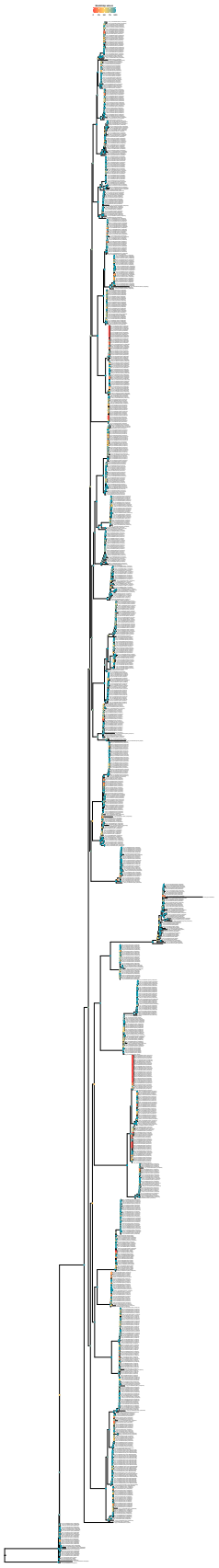
Supplementary Figure S3.4. The DSFG complement in Mammalia and *Drosophila* spp. Presence/absence of genes in various species are indicated by filled circles. Numbers inside each circle specify genes with 2 or more copies. The shaded area highlights outgroup species, *G. gallus* (Aves) for mammals and *A. gambiae* (Culicidae) for fruit flies. The phylogenetic tree of analysed species, as inferred from literature, is shown on the left, while major taxonomic groups are reported on the right. All species are represented by genomic data. DSFG trees are shown on the bottom (full trees can be found in **Supp. Fig. S5**–**S7**). Full species names for both mammals and fruit flies, along with all assembly and taxonomic information, can be found in **Supp. Tab. S3.4** and **Supp. Tab. S5**, respectively. A.: Aves; Chirop.: Chiroptera; L.: Lagomorpha; M.: Monotremata; Me.: Metatheria; P.: Pholidota; Pe.: Perissodactyla; Prim.: Primates; Roden.: Rodentia; X.: Xenarthra; C.: Culicidae.



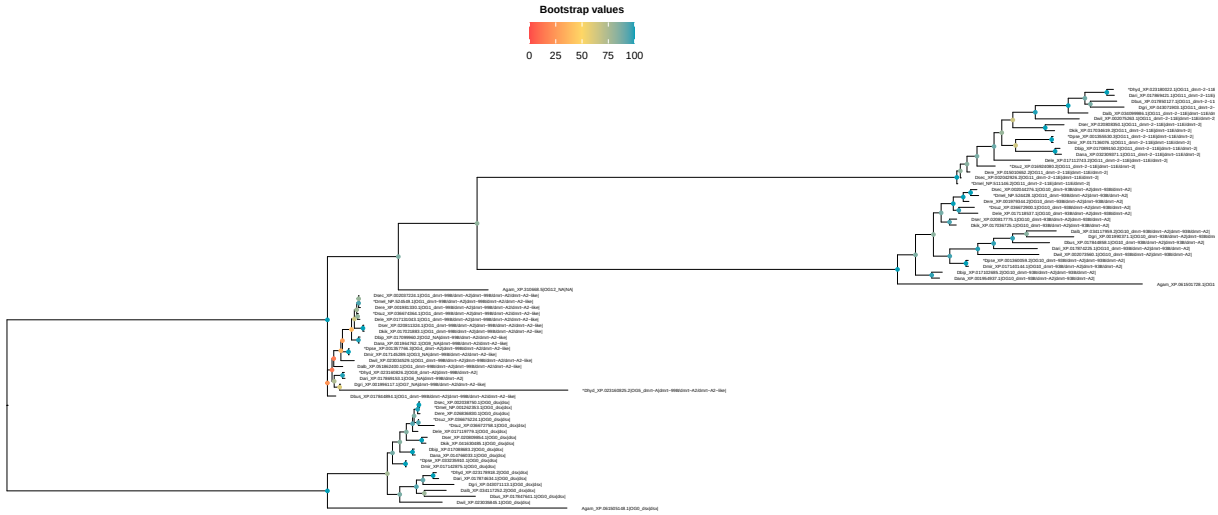
Supplementary Figure S3.5. ML phylogenetic tree of the Dmrt gene family in mammals, including the Possvm orthology inference. Reference genes from *H. sapiens*, *Mus musculus*, *Elephas maximus indicus*, and *Ornithorhynchus anatinus* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.4**. The tree has been midpoint rooted. Bootstrap values are shown for each node.



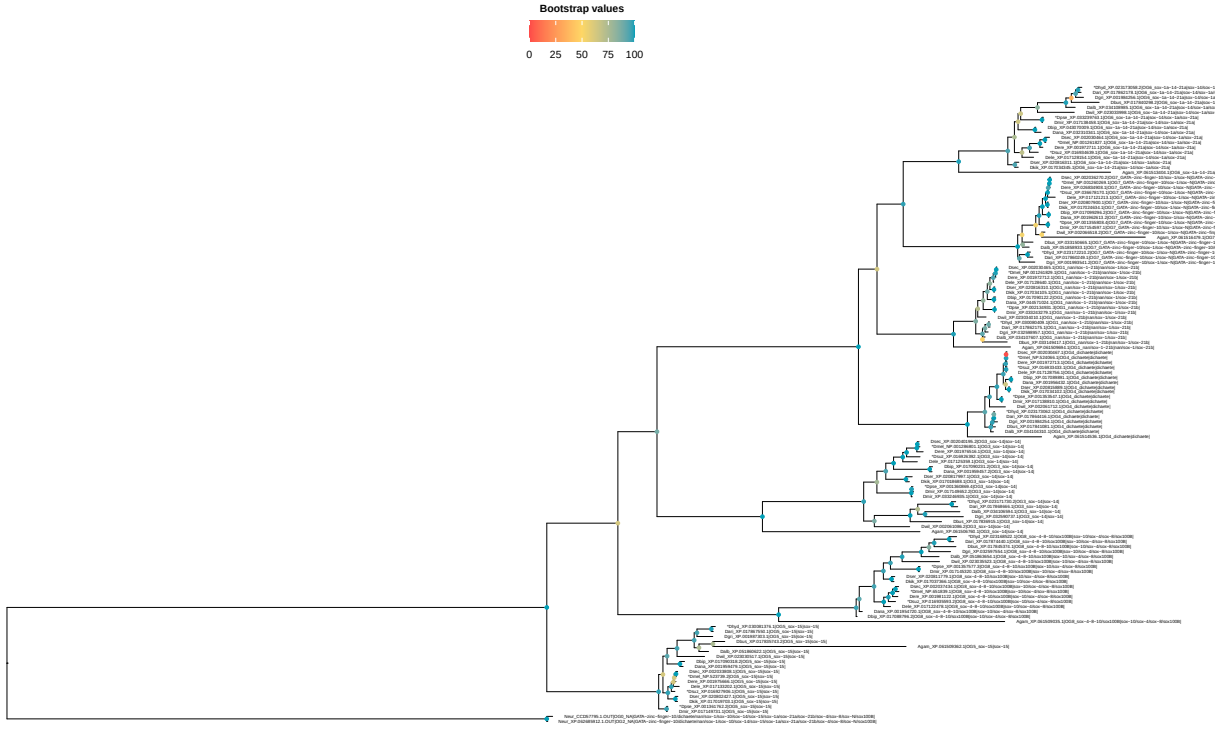
Supplementary Figure S3.6. ML phylogenetic tree of the Sox gene family in mammals, including the Possvm orthology inference. Reference genes from *H. sapiens*, *M. musculus*, *E. maximus indicus*, and *O. anatinus* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.4**. Bootstrap values are shown for each node.



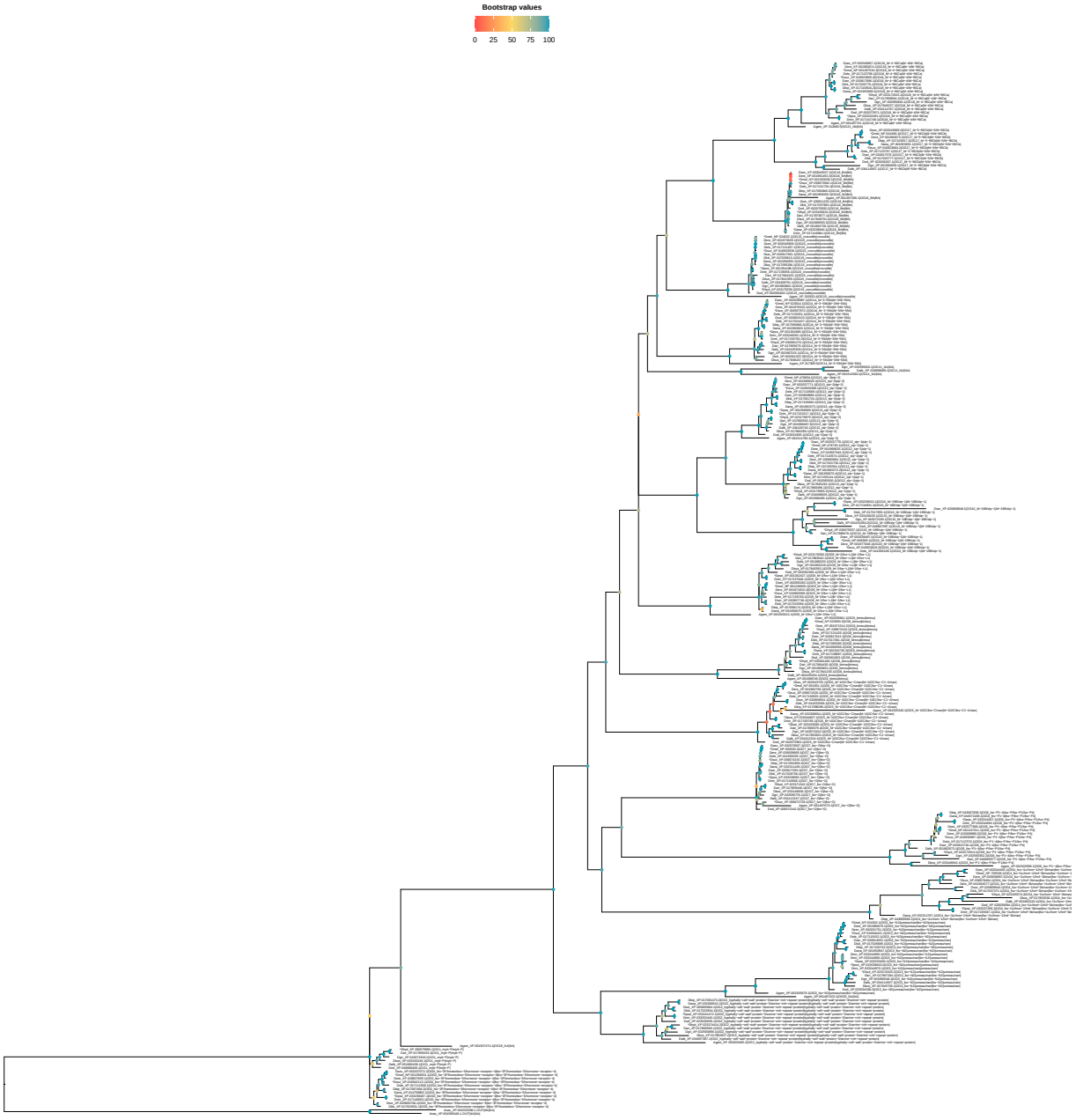
Supplementary Figure S3.7. ML phylogenetic tree of the Fox gene family in mammals, including the Possvm orthology inference. Reference genes from *H. sapiens*, *M. musculus*, *E. maximus indicus*, and *O. anatinus* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.4**. Bootstrap values are shown for each node.



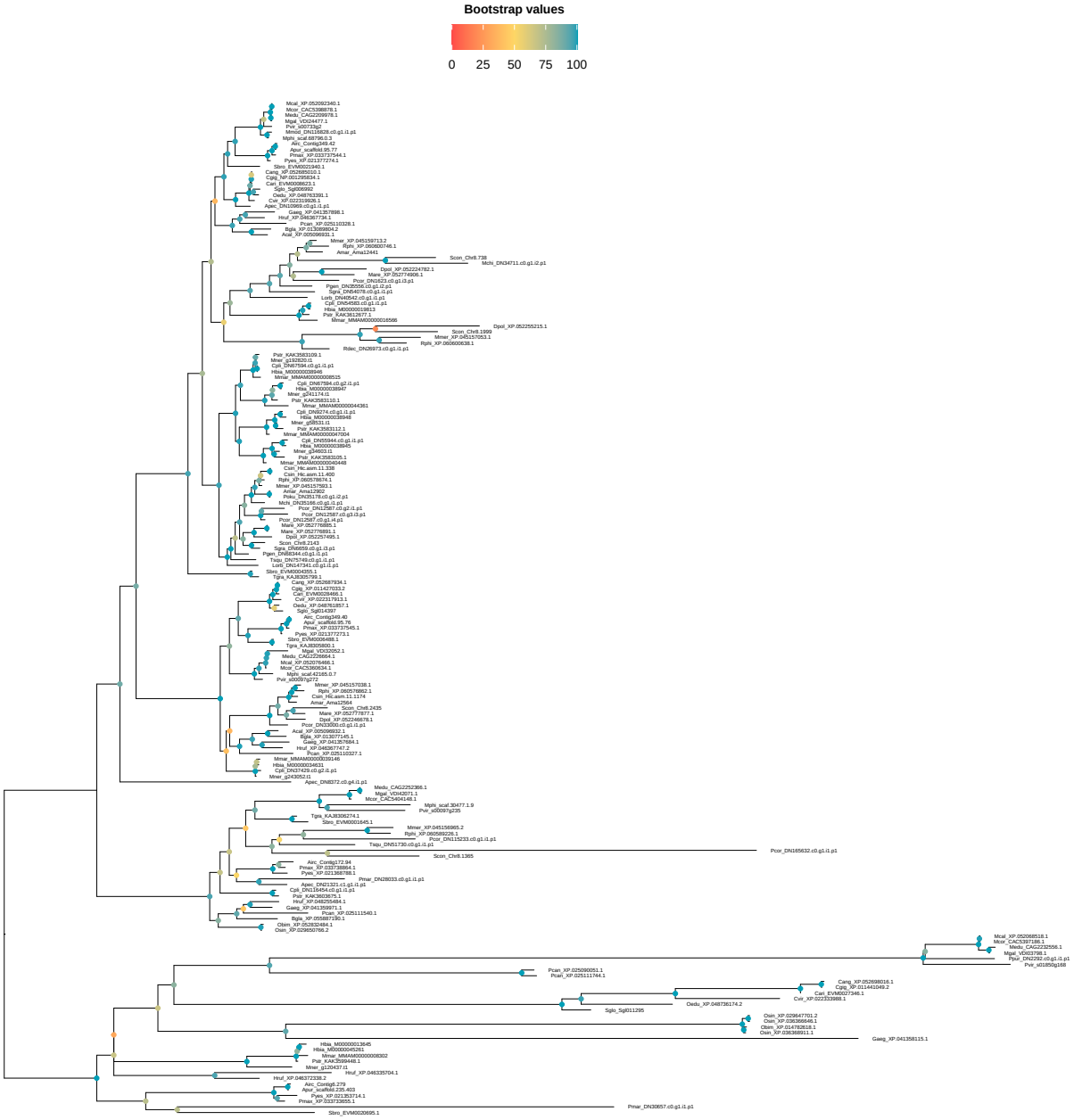
Supplementary Figure S3.8. ML phylogenetic tree of the Dmrt gene family in fruit flies, including the Possvm orthology inference. Reference genes from *D. melanogaster*, *Drosophila hydei*, *Drosophila pseudoobscura*, and *Drosophila suzukii* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S5**. The tree has been midpoint rooted. Bootstrap values are shown for each node.



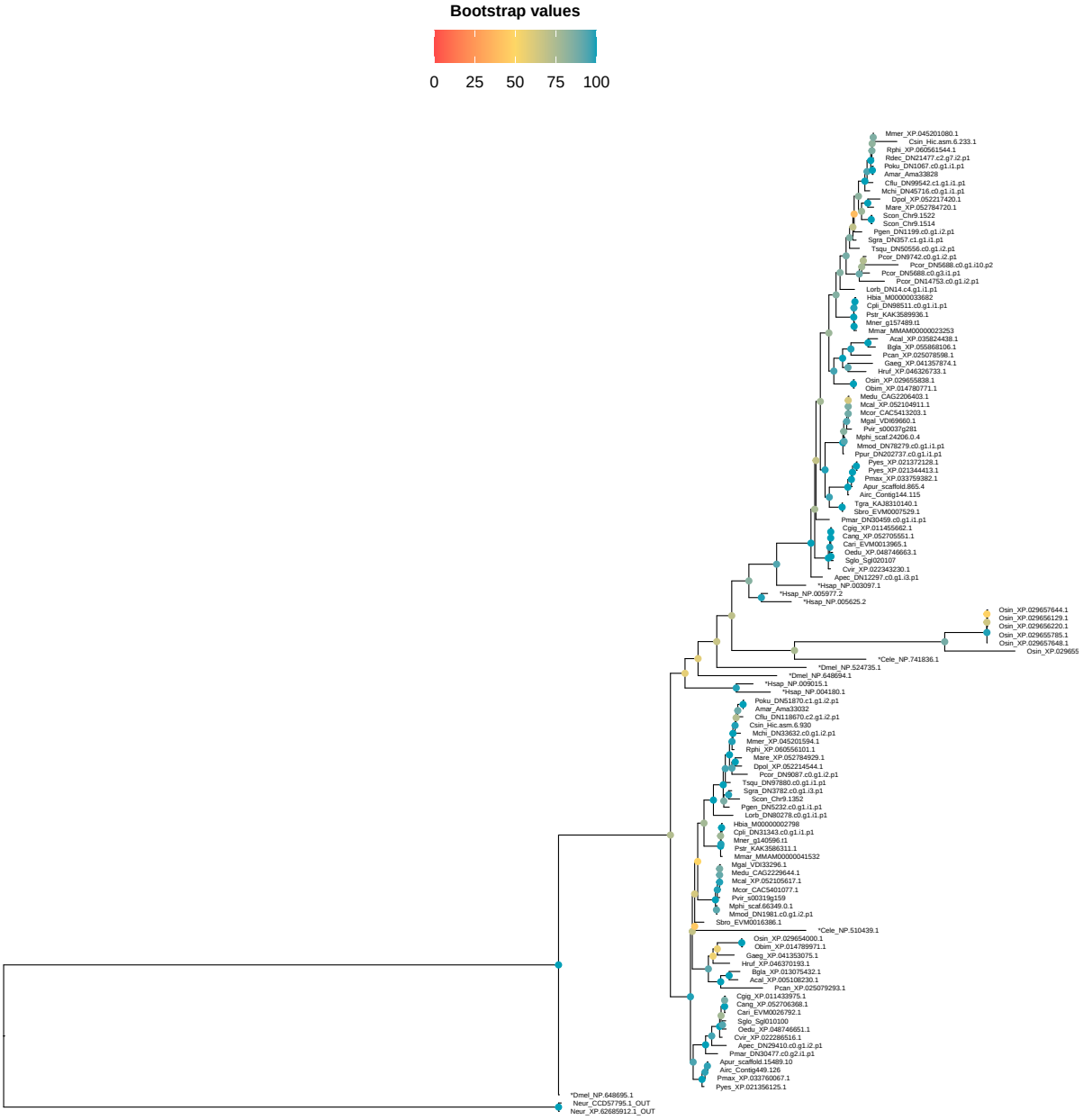
Supplementary Figure S3.9. ML phylogenetic tree of the Sox gene family in fruit flies, including the Possvm orthology inference. Reference genes from *D. melanogaster*, *D. hydei*, *D. pseudoobscura*, and *D. suzukii* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S5**. Bootstrap values are shown for each node.



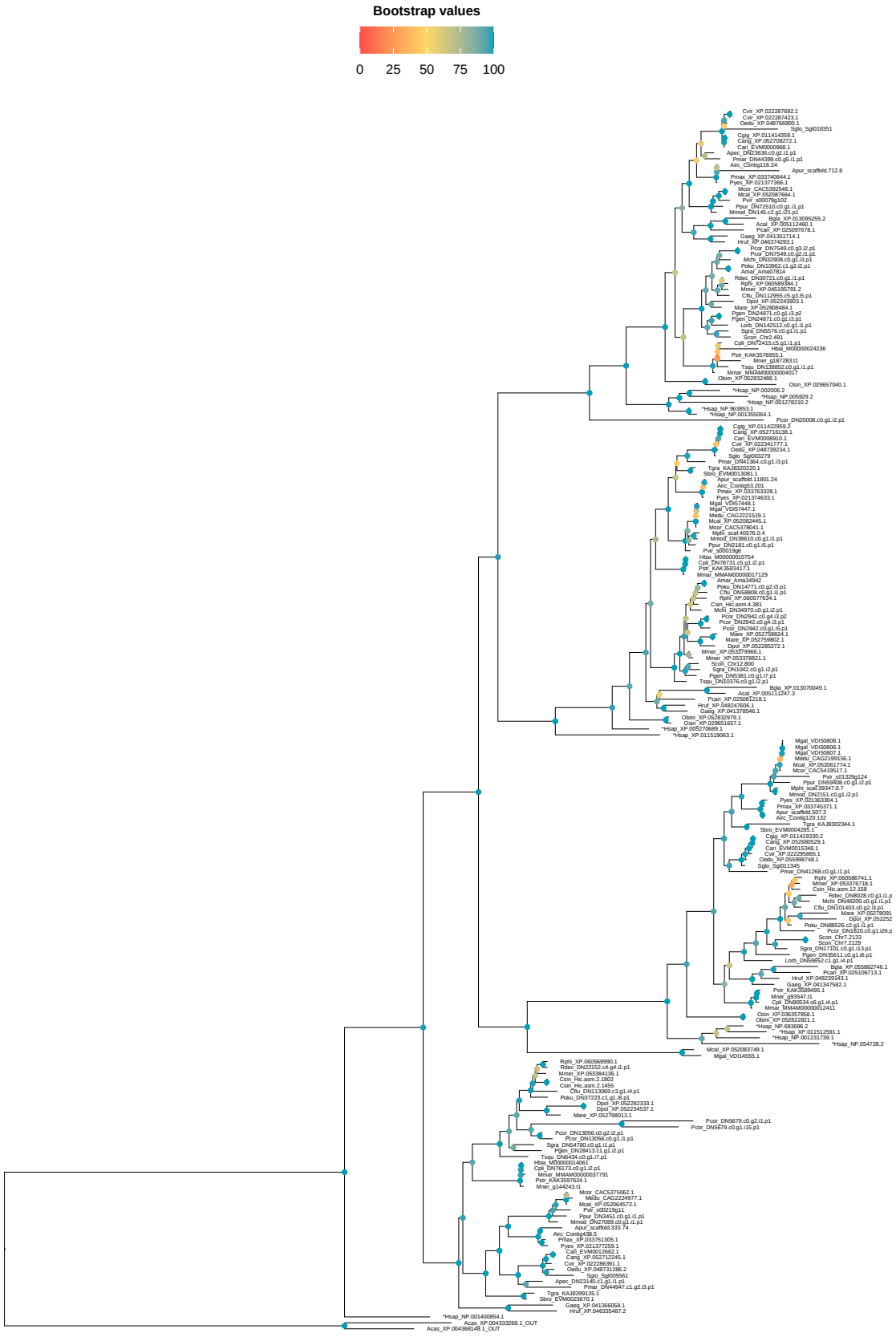
Supplementary Figure S3.10. ML phylogenetic tree of the Fox gene family in fruit flies, including the Possvm orthology inference. Reference genes from *D. melanogaster*, *D. hydei*, *D. pseudoobscura*, and *D. suzukii* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S5**. Bootstrap values are shown for each node.



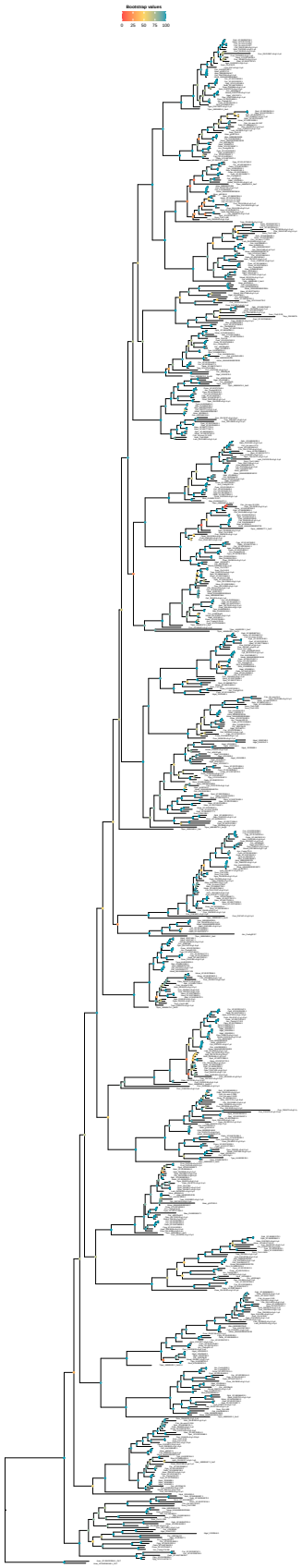
Supplementary Figure S3.11. ML phylogenetic tree of the Dmrt gene family in mollusc species. Species ID can be found in **Supp. Tab. S3.1**. The tree has been midpoint rooted. Bootstrap values are shown for each node.



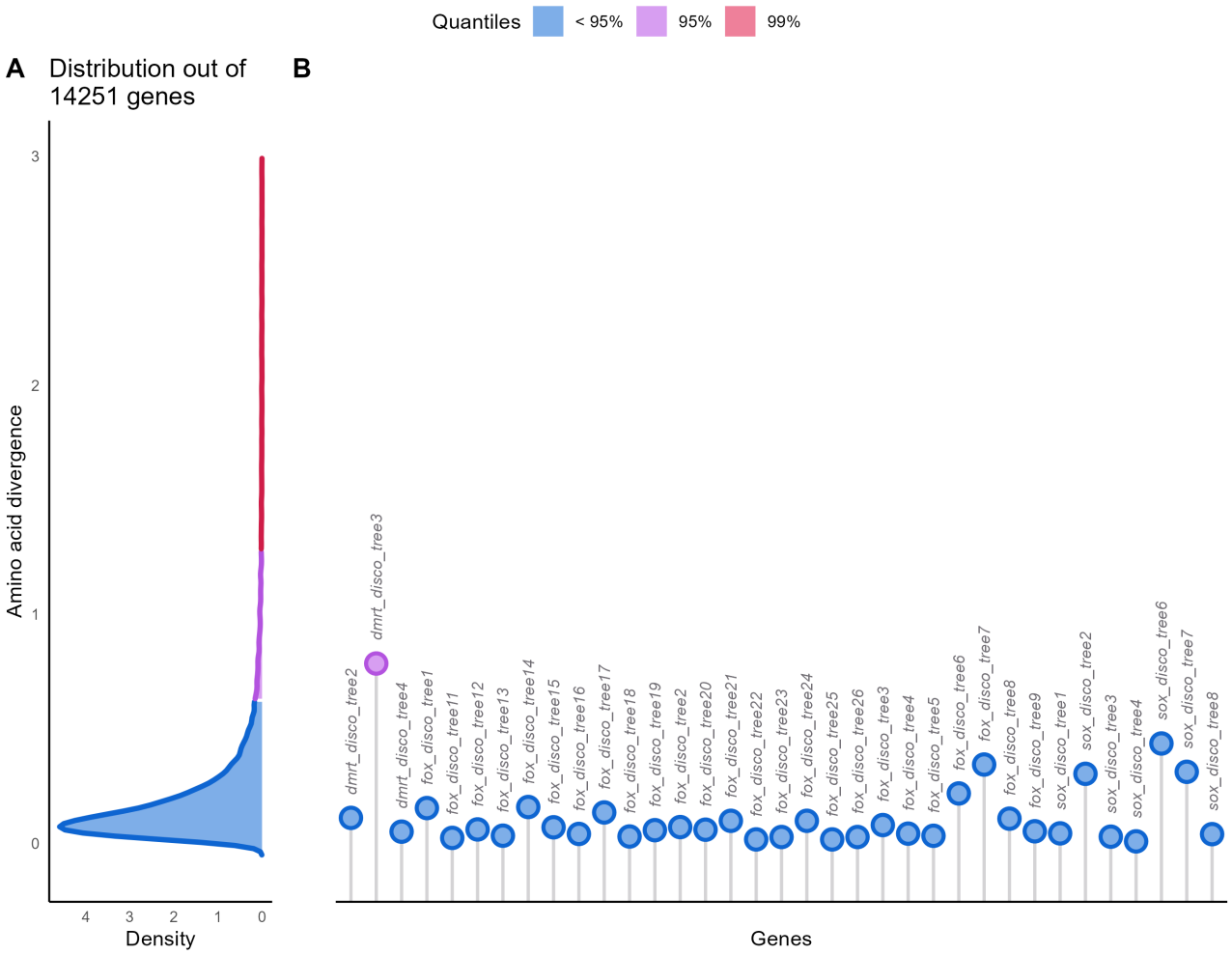
Supplementary Figure S3.12. ML phylogenetic tree of *Sox-B1* and *Sox-B2* genes in mollusc and reference species. Reference genes from *H. sapiens*, *C. elegans*, and *D. melanogaster* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.1**. Bootstrap values are shown for each node.



Supplementary Figure S3.13. ML phylogenetic tree of *Fox-J2*, *Fox-M*, *Fox-O*, and *Fox-P* genes in mollusc and reference species. Reference genes from *H. sapiens*, *C. elegans*, and *D. melanogaster* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.1**. Bootstrap values are shown for each node.



Supplementary Figure S3.14. ML phylogenetic tree of the Fox gene family in bivalves and the sea urchin *Strongylocentrotus purpuratus* (Spur). Reference genes from *S. purpuratus* are marked with an asterisk at the beginning of the tip names. Species ID can be found in **Supp. Tab. S3.1**. *S. purpuratus* genes are those given by **Tu et al., 2006**. Bootstrap values are shown for each node.



Supplementary Figure S3.15. Distribution of AASD of single-copy orthogroups in *Crassostrea gigas*, *Crassostrea angulata*, *Crassostrea ariakensis*, and *Crassostrea virginica* (A), including DSFG (B). The distribution of AASD in *Crassostrea* has been computed on the median values of pairwise distances of over 14k SCOs. Circle heights of DSFGs show the median value of their AASD. *Dmrt-1L* genes are indicated as 'dmrt_disco_tree3'.

3.6.2 Supplementary Tables

Supplementary Table S3.1. Genomic and transcriptomic data of bivalves and other molluscs.

Gene family	PANTHER/CDD	ID	Description

Supplementary Table S3.2. DSFG family and domain identifiers (IDs) in PANTHER and CDD, respectively. After having retrieved putative DSFGs on the basis of hidden Markov model (HMM) profiles, IDs have been used to retain only reliable hits.

Gene family	PANTHER/CDD	ID	Description
	CDD	gnl—CDD—214606	Doublesex DNA-binding motif
	CDD	gnl—CDD—425850	DM DNA binding domain
	PANTHER	PTHR12322	DOUBLESEX AND MAB-3 RELATED TRANSCRIPTION FACTOR DMRT
	PANTHER	PTHR12322:SF115	PROTEIN CBR-MAB-23
	PANTHER	PTHR12322:SF116	DOUBLESEX- AND MAB-3-RELATED TRANSCRIPTION FACTOR 1
Dmrt	PANTHER	PTHR12322:SF118	DOUBLESEX- AND MAB-3-RELATED TRANSCRIPTION FACTOR DMD-4
	PANTHER	PTHR12322:SF123	DOUBLESEX-MAB RELATED 99B
	PANTHER	PTHR12322:SF53	DOUBLESEX- AND MAB-3-RELATED TRANSCRIPTION FACTOR 2
	PANTHER	PTHR12322:SF71	DOUBLESEX- AND MAB-3-RELATED TRANSCRIPTION FACTOR A1
	PANTHER	PTHR16897:SF2	STRESS RESPONSE PROTEIN NST1
	PANTHER	PTHR46888:SF11	RIBONUCLEASE H
	CDD	gnl—CDD—432488	SOX transcription factor
	CDD	gnl—CDD—432558	Sox developmental protein N terminal
	CDD	gnl—CDD—438790	high mobility group (HMG)-box found in group B SRY-related high-mobility group (HMG) box (Sox) transcription factors
	CDD	gnl—CDD—438820	high mobility group (HMG)-box found in sex-determining region Y (SRY)-box (SOX) family transcription factors
	CDD	gnl—CDD—438837	high mobility group (HMG)-box found in group A, group B and group G of SRY-related high-mobility group (HMG) box (Sox) transcription factors
	CDD	gnl—CDD—438838	high mobility group (HMG)-box found in group C SRY-related high-mobility group (HMG) box (Sox) transcription factors
	CDD	gnl—CDD—438839	high mobility group (HMG)-box found in group D SRY-related high-mobility group (HMG) box (Sox) transcription factors
	CDD	gnl—CDD—438840	high mobility group (HMG)-box found in group E SRY-related high-mobility group (HMG) box (Sox) transcription factors
	CDD	gnl—CDD—438841	high mobility group (HMG)-box found in group F SRY-related high-mobility group (HMG) box (Sox) transcription factors
	CDD	gnl—CDD—438842	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 30 (SOX30) and similar proteins
	CDD	gnl—CDD—438843	high mobility group (HMG)-box found in sex-determining region Y protein (SRY) and similar proteins
	CDD	gnl—CDD—438844	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 15 (SOX15) and similar proteins
	CDD	gnl—CDD—438845	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 4 (SOX4) and similar proteins
	CDD	gnl—CDD—438846	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 11 (SOX11) and similar proteins
Sox	CDD	gnl—CDD—438847	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 12 (SOX12) and similar proteins
	CDD	gnl—CDD—438849	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 7 (SOX7) and similar proteins
	CDD	gnl—CDD—438850	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 17 (SOX17) and similar proteins
	CDD	gnl—CDD—438851	high mobility group (HMG)-box found in sex determining region Y (SRY)-box 18 (SOX18) and similar proteins
	PANTHER	PTHR10270:SF107	TRANSCRIPTION FACTOR SOX-14
	PANTHER	PTHR10270:SF161	SOX DOMAIN-CONTAINING PROTEIN DICHAETE-RELATED
	PANTHER	PTHR10270:SF199	SEX-DETERMINING REGION Y PROTEIN
	PANTHER	PTHR10270:SF231	TRANSCRIPTION FACTOR SOX-2
	PANTHER	PTHR10270:SF27	TRANSCRIPTION FACTOR SOX-4
	PANTHER	PTHR10270:SF313	TRANSCRIPTION FACTOR SOX-21
	PANTHER	PTHR10270:SF315	TRANSCRIPTION FACTOR SOX-1A-RELATED
	PANTHER	PTHR10270:SF317	TRANSCRIPTION FACTOR SOX-15-RELATED
	PANTHER	PTHR10270:SF322	TRANSCRIPTION FACTOR SOX-3
	PANTHER	PTHR10270:SF324	TRANSCRIPTION FACTOR SOX-3
	PANTHER	PTHR10270:SF326	TRANSCRIPTION FACTOR SOX

Table S3.2 continued from previous page

Gene family	PANTHER/CDD	ID	Description
	PANTHER	PTHR10270	SOX TRANSCRIPTION FACTOR
	PANTHER	PTHR45789	FI18025P1
	PANTHER	PTHR45789:SF2	FI18025P1
	PANTHER	PTHR45803:SF1	TRANSCRIPTION FACTOR SOX-9
Sox	PANTHER	PTHR45803:SF2	TRANSCRIPTION FACTOR SOX-8
	PANTHER	PTHR45803:SF5	SOX100B
	PANTHER	PTHR45803	SOX100B
	PANTHER	PTHR47279:SF1	TRANSCRIPTION FACTOR SOX-30
	PANTHER	PTHR47279	TRANSCRIPTION FACTOR SOX-30
	CDD	gnl—CDD—410788	Forkhead (FH) domain found in Forkhead box (FOX) family of transcription factors and similar proteins
	CDD	gnl—CDD—410789	Forkhead (FH) domain found in the Forkhead box protein A (FOXA) subfamily
	CDD	gnl—CDD—410790	Forkhead (FH) domain found in the Forkhead box protein B (FOXB) subfamily
	CDD	gnl—CDD—410791	Forkhead (FH) domain found in the Forkhead box protein C (FOXC) subfamily
	CDD	gnl—CDD—410792	Forkhead (FH) domain found in the Forkhead box protein D (FOXD) subfamily
	CDD	gnl—CDD—410793	Forkhead (FH) domain found in the Forkhead box protein E (FOXE) subfamily
	CDD	gnl—CDD—410794	Forkhead (FH) domain found in the Forkhead box protein F (FOXF) subfamily
	CDD	gnl—CDD—410795	Forkhead (FH) domain found in the Forkhead box protein G (FOXG) subfamily
	CDD	gnl—CDD—410796	Forkhead (FH) domain found in the Forkhead box protein H (FOXH) subfamily
	CDD	gnl—CDD—410797	Forkhead (FH) domain found in Forkhead box protein J1 (FOXJ1) and similar proteins
	CDD	gnl—CDD—410798	Forkhead (FH) domain found in Forkhead box proteins, FOXJ2, FOXJ3 and similar proteins
	CDD	gnl—CDD—410799	Forkhead (FH) domain found in the Forkhead box protein I (FOXI) subfamily
	CDD	gnl—CDD—410800	Forkhead (FH) domain found in the Forkhead box protein K (FOXK) subfamily
	CDD	gnl—CDD—410801	Forkhead (FH) domain found in Forkhead box protein L1 (FOXL1) and similar proteins
	CDD	gnl—CDD—410802	Forkhead (FH) domain found in Forkhead box protein L2 (FOXL2) and similar proteins
	CDD	gnl—CDD—410803	Forkhead (FH) domain found in the Forkhead box protein M (FOXM) subfamily
Fox	CDD	gnl—CDD—410804	Forkhead (FH) domain found in Forkhead box protein N1 (FOXN1) and similar proteins
	CDD	gnl—CDD—410805	Forkhead (FH) domain found in Forkhead box protein N2 (FOXN2) and similar proteins
	CDD	gnl—CDD—410806	Forkhead (FH) domain found in the Forkhead box protein O (FOXO) subfamily
	CDD	gnl—CDD—410807	Forkhead (FH) domain found in the Forkhead box protein P (FOXP) subfamily
	CDD	gnl—CDD—410808	Forkhead (FH) domain found in Forkhead box protein Q1 (FOXQ1) and similar proteins
	CDD	gnl—CDD—410809	Forkhead (FH) domain found in Forkhead box protein Q2 (FOXQ2) and similar proteins
	CDD	gnl—CDD—410810	Forkhead (FH) domain found in the Forkhead box protein R (FOXR) subfamily
	CDD	gnl—CDD—410811	Forkhead (FH) domain found in Forkhead box protein S1 (FOXS1)
	CDD	gnl—CDD—410812	Forkhead (FH) domain found in Forkhead box protein A1 (FOXA1) and similar proteins
	CDD	gnl—CDD—410813	Forkhead (FH) domain found in Forkhead box protein A2 (FOXA2) and similar proteins
	CDD	gnl—CDD—410814	Forkhead (FH) domain found in Forkhead box protein A3 (FOXA3) and similar proteins
	CDD	gnl—CDD—410816	Forkhead (FH) domain found in Forkhead box protein B1 (FOXB1) and similar proteins
	CDD	gnl—CDD—410817	Forkhead (FH) domain found in Forkhead box protein B2 (FOXB2) and similar proteins
	CDD	gnl—CDD—410818	Forkhead (FH) domain found in Forkhead box protein C1 (FOXC1) and similar proteins
	CDD	gnl—CDD—410819	Forkhead (FH) domain found in Forkhead box protein C2 (FOXC2) and similar proteins
	CDD	gnl—CDD—410820	Forkhead (FH) domain found in Forkhead box proteins FOXD1, FOXD2 and similar proteins
	CDD	gnl—CDD—410821	Forkhead (FH) domain found in Forkhead box protein D3 (FOXD3) and similar proteins

Table S3.2 continued from previous page

Gene family	PANTHER/CDD	ID	Description
	CDD	gnl—CDD—410822	Forkhead (FH) domain found in Forkhead box protein D4 (FOXD4) and similar proteins
	CDD	gnl—CDD—410823	Forkhead (FH) domain found in Forkhead box protein F1 (FOXF1) and similar proteins
	CDD	gnl—CDD—410824	Forkhead (FH) domain found in Forkhead box protein F2 (FOXF2) and similar proteins
	CDD	gnl—CDD—410825	Forkhead (FH) domain found in Forkhead box protein J2 (FOXJ2) and similar proteins
	CDD	gnl—CDD—410826	Forkhead (FH) domain found in Forkhead box protein J3 (FOXJ3) and similar proteins
	CDD	gnl—CDD—410827	Forkhead (FH) domain found in Forkhead box protein I1 (FOXI1) and similar proteins
	CDD	gnl—CDD—410828	Forkhead (FH) domain found in Forkhead box protein K1 (FOXK1) and similar proteins
	CDD	gnl—CDD—410829	Forkhead (FH) domain found in Forkhead box protein K2 (FOXK2) and similar proteins
	CDD	gnl—CDD—410830	Forkhead (FH) domain found in Forkhead box protein N1 (FOXN1)
	CDD	gnl—CDD—410831	Forkhead (FH) domain found in Forkhead box protein N4 (FOXN4)
	CDD	gnl—CDD—410832	Forkhead (FH) domain found in Forkhead box protein N2 (FOXN2)
	CDD	gnl—CDD—410833	Forkhead (FH) domain found in Forkhead box protein N3 (FOXN3)
	CDD	gnl—CDD—410834	Forkhead (FH) domain found in Forkhead box protein O1 (FOXO1)
	CDD	gnl—CDD—410835	Forkhead (FH) domain found in Forkhead box protein O3 (FOXO3)
	CDD	gnl—CDD—410836	Forkhead (FH) domain found in Forkhead box protein O4 (FOXO4) and similar proteins
	CDD	gnl—CDD—410837	Forkhead (FH) domain found in Forkhead box protein O6 (FOXO6) and similar proteins
	CDD	gnl—CDD—410838	Forkhead (FH) domain found in Forkhead box protein P1 (FOXP1)
	CDD	gnl—CDD—410839	Forkhead (FH) domain found in Forkhead box protein P2 (FOXP2)
	CDD	gnl—CDD—410840	Forkhead (FH) domain found in Forkhead box protein P3 (FOXP3) and similar proteins
	CDD	gnl—CDD—410841	Forkhead (FH) domain found in Forkhead box protein P4 (FOXP4) and similar proteins
	PANTHER	PTHR11829	FORKHEAD BOX PROTEIN
Fox	PANTHER	PTHR11829:SF142	FOXQ2 PROTEIN
	PANTHER	PTHR11829:SF156	FORKHEAD BOX PROTEIN E3
	PANTHER	PTHR11829:SF206	FORKHEAD BOX PROTEIN Q1
	PANTHER	PTHR11829:SF209	FORKHEAD BOX PROTEIN B1
	PANTHER	PTHR11829:SF335	FORKHEAD BOX PROTEIN D2
	PANTHER	PTHR11829:SF340	FORKHEAD BOX PROTEIN H1
	PANTHER	PTHR11829:SF342	FORKHEAD BOX PROTEIN L2
	PANTHER	PTHR11829:SF348	FORKHEAD BOX PROTEIN D1
	PANTHER	PTHR11829:SF361	FORKHEAD BOX PROTEIN D3
	PANTHER	PTHR11829:SF398	FORKHEAD BOX PROTEIN PES-1
	PANTHER	PTHR11829:SF399	FORKHEAD TRANSCRIPTION FACTOR FKH-9
	PANTHER	PTHR11829:SF401	FORKHEAD BOX C1-A-RELATED
	PANTHER	PTHR13962	FORKHEAD BOX PROTEIN N3-LIKE PROTEIN-RELATED
	PANTHER	PTHR13962:SF17	FORKHEAD BOX PROTEIN N4
	PANTHER	PTHR13962:SF19	FORKHEAD BOX PROTEIN N2
	PANTHER	PTHR13962:SF20	FORKHEAD BOX PROTEIN N3
	PANTHER	PTHR13962:SF22	FORKHEAD BOX PROTEIN N3-LIKE PROTEIN
	PANTHER	PTHR13962:SF26	FORKHEAD BOX PROTEIN N2
	PANTHER	PTHR45767	FORKHEAD BOX PROTEIN O
	PANTHER	PTHR45767:SF2	FORKHEAD BOX PROTEIN O
	PANTHER	PTHR45796	FORKHEAD BOX P, ISOFORM C
	PANTHER	PTHR45796:SF3	FORKHEAD BOX PROTEIN P1

Table S3.2 continued from previous page

Gene family	PANTHER/CDD	ID	Description
	PANTHER	PTHR45796:SF4	FORKHEAD BOX P, ISOFORM C
	PANTHER	PTHR45881:SF3	FORKHEAD BOX PROTEIN K2
	PANTHER	PTHR45881:SF4	FORKHEAD BOX PROTEIN K1
	PANTHER	PTHR46078	FORKHEAD BOX PROTEIN J2 FAMILY MEMBER
	PANTHER	PTHR46262	FORKHEAD BOX PROTEIN BINIOU
	PANTHER	PTHR46262:SF2	FORKHEAD BOX PROTEIN BINIOU
	PANTHER	PTHR46617	FORKHEAD BOX PROTEIN G1
Fox	PANTHER	PTHR46617:SF3	FORKHEAD BOX PROTEIN G1
	PANTHER	PTHR46721	FORKHEAD BOX PROTEIN N1
	PANTHER	PTHR46721:SF2	FORKHEAD BOX N1
	PANTHER	PTHR46805	FORKHEAD BOX PROTEIN J1
	PANTHER	PTHR46878	FORKHEAD BOX PROTEIN M1
	PANTHER	PTHR46878:SF1	FORKHEAD BOX PROTEIN M1
	PANTHER	PTHR47316	FORKHEAD BOX PROTEIN H1
	PANTHER	PTHR47316:SF1	FORKHEAD BOX PROTEIN H1

Supplementary Table S3.3. List of DSFGs from reference species used to assess the identity of DSFGs in molluscs. NCBI accession numbers are reported in parenthesis. Each row represents an orthology group.

Homo sapiens	*Drosophila melanogaster*	*Caenorhabditis elegans*	Group
Dmrt gene family			
DMRT1 (NP_068770.2)	-	-	1
DMRT2 (NP_006548.1)	*dmrt11E (NP_511146.2)*	-	2
DMRT3 (NP_067063.1)	*dmrt93B (NP_524428.1)*	*dmd-4 (NP_510466.1)*	3
DMRT4/A1 (NP_071443.2) *DMRT5/A2 (NP_115486.1)*	*dmrt99b (NP_524549.1)*	*dmd-5 (NP_495138.2)*	A1/2
DMRT6/B1 (NP_149056.1)	-	-	-
DMRT7/C2 (NP_001035373.1)	-	-	-
DMRT8/C1 (NP_149042.2)	-	-	-
-	*dsx (NP_731197.1)*	-	-
-	-	*mab3 (NP_001256882.1)*	-
-	-	*dmd-3 (NP_001256883.1)*	-
-	-	*dmd-6 (NP_001370045.1)*	-
-	-	*dmd-7 (NP_741551.1)*	-
-	-	*dmd-8 (NP_503176.2)*	-
-	-	*dmd-9 (NP_500305.1)*	-
-	-	*dmd-11 (NP_001379162.1)*	-
-	-	*mab-23 (NP_001041089.1)*	-
Sox gene family			
SRY (NP_003131.1)	-	-	A
SOX3 (NP_005625.2) *SOX2 (NP_003097.1)* *SOX1 (NP_005977.2)*	dichaete (NP_524066.1) soxN (NP_524735.1)	sox3 (NP_510439.1) sox2 (NP_741836.1)	B1
SOX14 (NP_004180.1) *SOX21 (NP_009015.1)*	sox21a (NP_648694.1) sox21b (NP_648695.1)	-	B2
SOX11 (NP_003099.1) *SOX12 (NP_008874.2)* *SOX4 (NP_003098.1)*	*sox14 (NP_476894.1)*	*sem-2 (NP_740846.1)*	C
SOX13 (NP_005677.2) *SOX5 (NP_008871.3)* *SOX6 (NP_001139291.2)*	*sox102f (NP_726612.1)*	*egl-13 (NP_001024918.1)*	D
SOX9 (NP_000337.1) *SOX8 (NP_055402.2)* *SOX10 (NP_008872.1)*	*sox110b (NP_651839.1)*	-	E
SOX18 (NP_060889.1) *SOX7 (NP_113627.1)* *SOX17 (NP_071899.1)*	*sox15 (NP_523739.2)*	-	F
SOX15 (NP_008873.1)	-	-	G
SOX30 (NP_848511.1)	-	-	H
Fox gene family			
FOXA1/HNF-3α (NP_004487.2) *FOXA2/HNF-3β (NP_068556.2)* *FOXA3/HNF-3γ (NP_004488.2)*	*forkhead/fkh (NP_524542.1)*	*pha-4/Ce-fkh1 (NP_001041114.1)*	A
FOXB1 (NP_036314.2) *FOXB2 (NP_001013757.1)*	*fd96Ca/fd4 (NP_524495.1)* *fd96Cb/fd5 (NP_524496.1)*	*lin-31 (NP_494704.1)*	B
FOXC1/MF1/FKHL7 (NP_001444.2) *FOXC2/MFH1 (NP_005242.1)*	*crocodile/fd1 (NP_524202.1)*	-	C

Table S3.3 continued from previous page

Homo sapiens	*Drosophila melanogaster*	*Caenorhabditis elegans*	Group
	Fox gene family		
FOXD1/FREAC4 (NP_004463.1) *FOXD2/FREAC9 (NP_004465.3)* *FOXD3 (NP_036315.1)* *FOXD4 (NP_997188.2)*	*fd59A/fd3 (NP_523814.1)*	*unc-130 (NP_496411.1)*	D
FOXE1/TITF2 (NP_004464.2) *FOXE3 (NP_036318.1)*	-	-	E
FOXF1 (NP_001442.2) *FOXF2 (NP_001443.1)*	*binious/FoxF (NP_523950.2)*	*let-381/F26B1.7 (NP_491826.1)*	F
FOXG1/BF1/HBF2 (NP_005240.3)	*slp1 (NP_476730.1)* *slp2 (NP_476834.1)* *fd19B/cg9571 (NP_608369.1)*	*fkh2/T14G12.4 (NP_508644.1)*	G
FOXH1/FAST1 (NP_003914.1)	-	-	H
FOXI1/FREAC6/HFH3 (NP_036320.2)	-	-	I
FOXJ1 (NP_001445.2)	-	-	J1
FOXJ2 (XP_011519063.1)	-	-	J2
FOXJ3 (XP_005270689.1)	-	-	J3
FOXK1/ILF1 (NP_001032242.1) *FOXK2 (NP_004505.2)*	*foxK/LD16137 (NP_001261701.1)*	-	K
FOXL1 (NP_005241.1)	*foxL1/fd2 (NP_523912.1)*	-	L1
FOXL2 (NP_075555.1)	-	-	L2
FOXM1 (NP_001400854.1)	-	-	M
FOXN1/WHN (NP_001356298.1) *FOXN4 (NP_998761.2)*	*jumeau (NP_524302.1)*	-	N1/4
FOXN2/HTLF (NP_001362376.1) *FOXN3/CHES1 (NP_001078940.1)*	*ches-1 (NP_511071.3)*	-	N2/3
FOXO1 (NP_002006.2) *FOXO3 (NP_963853.1)* *FOXO3B (NP_001355064.1)*	-	*daf-16 (NP_001364785.1)*	O
FOXP1 (NP_001231739.1) *FOXP2 (NP_683696.2)* *FOXP3 (NP_054728.2)* *FOXP4 (XP_011512591.1)*	*foxP/cg16899 (NP_001247011.1)*	*F26D12.1 (NP_001293813.1)*	P
FOXQ/HFH11 (NP_150285.3)	-	-	Q1
-	*fd102C/cd11152 (NP_651951.1)*	*fkh-10/C25A1.2 (NP_492676.2)*	Q2
FOXS1/FREAC10 (NP_004109.1)	-	-	S
-	-	*PES-1 (NP_001023406.1)*	-
-	-	*B0286.5/FKH-6 (NP_494775.1)*	-
-	-	*F40H3.4/FKH-8 (NP_001254107.1)*	-
-	-	*C29F7.4/FKH-3 (NP_001294822.1)*	-
-	-	*K03C7.2/FKH-9 (NP_001024760.1)*	-

Supplementary Table S3.4. Genomic data of mammals used to retrieve DSFGs and compute AASD of SCOs. For each species, the relative ID, taxonomic information, BUSCO statistics, NCBI accession number, and source publication. are reported.

Species	ID	Class	Group	Order	Type	BUSCO statistics (mammalia_odb10)	NCBI acc. no.	Reference
Gallus gallus	Ggal	Aves	Neognathae	Galliformes	Genome	C:99.0%[S:98.6%,D:0.4%],F:0.2%,M:0.8%	GCF_016699485.2	**Vertebrate Genome Project**
Chrysochloris asiatica	Casi	Mammalia	Afrotheria	Afroscoricida	Genome	C:98.0%[S:97.4%,D:0.6%],F:1.1%,M:0.9%	GCF_000296735.1	**Murata et al., 2003**
Elephas maximus indicus	Emax	Mammalia	Afrotheria	Proboscidea	Genome	C:98.9%[S:98.3%,D:0.6%],F:0.4%,M:0.7%	GCF_024166365.1	**Vertebrate Genome Project**
Trichechus manatus latirostris	Tman	Mammalia	Afrotheria	Sirenia	Genome	C:96.1%[S:95.7%,D:0.4%],F:1.8%,M:2.1%	GCF_000243295.1	**Foote et al., 2015**
Orycteropus afer afer	Oafe	Mammalia	Afrotheria	Tubulidentata	Genome	C:96.5%[S:96.0%,D:0.5%],F:1.9%,M:1.6%	GCF_000298275.1	N/A
Ochotona princeps	Opri	Mammalia	Euarchontoglires	Lagomorpha	Genome	C:98.3%[S:96.4%,D:1.9%],F:0.5%,M:1.2%	GCF_030435755.1	**Vertebrate Genome Project**
Cebus imitator	Cimi	Mammalia	Euarchontoglires	Primates	Genome	C:97.3%[S:95.1%,D:2.2%],F:1.7%,M:1.0%	GCF_001604975.1	**Orkin et al., 2021**
Homo sapiens	Hsap	Mammalia	Euarchontoglires	Primates	Genome	C:99.6%[S:97.3%,D:2.3%],F:0.2%,M:0.2%	GCF_000001405.40	**Genome Reference Consortium**
Lemur catta	Lcat	Mammalia	Euarchontoglires	Primates	Genome	C:98.3%[S:97.2%,D:1.1%],F:0.4%,M:1.3%	GCF_020740605.2	**Vertebrate Genome Project**
Cavia porcellus	Cpor	Mammalia	Euarchontoglires	Rodentia	Genome	C:96.4%[S:95.7%,D:0.7%],F:1.7%,M:1.9%	GCF_000151735.1	**The Genome Sequencing Platform**
Mus musculus	Mmus	Mammalia	Euarchontoglires	Rodentia	Genome	C:99.4%[S:98.7%,D:0.7%],F:0.2%,M:0.4%	GCF_000001635.27	**Genome Reference Consortium**
Sciurus carolinensis	Scar	Mammalia	Euarchontoglires	Rodentia	Genome	C:99.1%[S:96.9%,D:2.2%],F:0.3%,M:0.6%	GCF_902686445.1	**Mead et al., 2020**
Bubalus bubalis	Bbub	Mammalia	Laurasiatheria	Artiodactyla	Genome	C:98.7%[S:97.0%,D:1.7%],F:0.6%,M:0.7%	GCF_019923935.1	**Deng et al., 2016**
Balaenoptera musculus	Bmus	Mammalia	Laurasiatheria	Artiodactyla	Genome	C:98.4%[S:95.7%,D:2.7%],F:0.6%,M:1.0%	GCF_009873245.2	**Genome 10K**
Camelus dromedarius	Cdro	Mammalia	Laurasiatheria	Artiodactyla	Genome	C:98.7%[S:98.3%,D:0.4%],F:0.7%,M:0.6%	GCF_000803125.2	**Elbers et al., 2019**
Hippopotamus amphibius kiboko	Hamp	Mammalia	Laurasiatheria	Artiodactyla	Genome	C:98.7%[S:95.2%,D:3.5%],F:0.5%,M:0.8%	GCF_030028045.1	**Vertebrate Genome Project**
Phacochoerus africanus	Pafr	Mammalia	Laurasiatheria	Artiodactyla	Genome	C:98.8%[S:98.3%,D:0.5%],F:0.6%,M:0.6%	GCF_016906955.1	N/A
Tursiops truncatus	Ttru	Mammalia	Laurasiatheria	Artiodactyla	Genome	C:97.3%[S:95.2%,D:2.1%],F:1.1%,M:1.6%	GCF_011762595.1	**Xiong et al., 2009**
Ailuropoda melanoleuca	Amel	Mammalia	Laurasiatheria	Carnivora	Genome	C:97.3%[S:96.6%,D:0.7%],F:1.3%,M:1.4%	GCF_002007445.2	**Fan et al., 2019**
Canis lupus familiaris	Clup	Mammalia	Laurasiatheria	Carnivora	Genome	C:98.5%[S:96.7%,D:1.8%],F:0.6%,M:0.9%	GCF_011100685.1	**Wang et al., 2021**
Mirounga angustirostris	Mang	Mammalia	Laurasiatheria	Carnivora	Genome	C:96.7%[S:94.5%,D:2.2%],F:1.9%,M:1.4%	GCF_021288785.2	**Moreno et al., 2024**
Panthera tigris	Ptig	Mammalia	Laurasiatheria	Carnivora	Genome	C:99.4%[S:98.9%,D:0.5%],F:0.3%,M:0.3%	GCF_018350195.1	**Bredemeyer et al., 2023**
Desmodus rotundus	Drot	Mammalia	Laurasiatheria	Chiroptera	Genome	C:98.2%[S:97.2%,D:1.0%],F:0.5%,M:1.3%	GCF_022682495.1	**Bat 1K**
Pteropus giganteus	Pgig	Mammalia	Laurasiatheria	Chiroptera	Genome	C:97.2%[S:96.9%,D:0.3%],F:1.1%,M:1.7%	GCF_902729225.1	**Fouret et al., 2020**
Rhinolophus ferrumequinum	Rfer	Mammalia	Laurasiatheria	Chiroptera	Genome	C:99.2%[S:97.9%,D:1.3%],F:0.3%,M:0.5%	GCF_004115265.2	**Vertebrate Genome Project**
Ceratotherium simum simum	Csim	Mammalia	Laurasiatheria	Perissodactyla	Genome	C:98.8%[S:98.6%,D:0.2%],F:0.9%,M:0.3%	GCF_000283155.1	N/A
Equus quagga	Equa	Mammalia	Laurasiatheria	Perissodactyla	Genome	C:98.5%[S:95.0%,D:3.5%],F:0.5%,M:1.0%	GCF_021613505.1	**Vilstrup et al., 2013**
Manis javanica	Mjav	Mammalia	Laurasiatheria	Pholidota	Genome	C:95.7%[S:93.7%,D:2.0%],F:1.9%,M:2.4%	GCF_014570535.1	N/A
Sarcophilus harrisii	Shar	Mammalia	Metatheria	Dasyuromorphia	Genome	C:95.5%[S:94.5%,D:1.0%],F:0.9%,M:3.6%	GCF_902635505.1	**Stammnitz et al., 2023**
Monodelphis domestica	Mdom	Mammalia	Metatheria	Didelphimorphia	Genome	C:95.1%[S:92.3%,D:2.8%],F:0.9%,M:4.0%	GCF_027887165.1	**Vertebrate Genome Project**
Ornithorhynchus anatinus	Oana	Mammalia	Prototheria	Monotremata	Genome	C:92.3%[S:91.2%,D:1.1%],F:1.4%,M:6.3%	GCF_004115215.2	**zhou2021platypus**
Dasypus novemcinctus	Dnov	Mammalia	Xenarthra	Cingulata	Genome	C:96.9%[S:94.3%,D:2.6%],F:0.7%,M:2.4%	GCF_030445035.1	**Vertebrate Genome Project**
Choloepus didactylus	Cdid	Mammalia	Xenarthra	Pilosa	Genome	C:97.8%[S:91.9%,D:5.9%],F:0.7%,M:1.5%	GCF_015220235.1	**Vertebrate Genome Project**

Chapter 4

Expression patterns of three sex-related genes and the germline marker *Vasa* in early developmental stages of *Mytilus galloprovincialis* embryos

Filippo Nicolini[1,2], Sergey Nuzhdin[3], Fabrizio Ghiselli[1], Andrea Luchetti[1], Liliana Milani[1]

[1] *Department of Biological, Geological and Environmental Science, University of Bologna, Bologna (BO), Italy.*

[2] *Fano Marine Center, Fano (PU), Italy.*

[3] *3Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA.*

In preparation.

4.1 Introduction

In preparation.

4.2 Materials and Methods

4.2.1 Time-series gene expression

Miglioli et al., 2024 recently produced one of the very first detailed developmental transcriptome of *M. galloprovincialis*, spanning from the unfertilized oocyte to the larval stage at 72 hpf, with time points sampled every 4 hpf. A total of 30 different mRNA libraries was sequenced, consisting of fifteen developmental time points per two technical replicates. These data are very useful to thoroughly investigate the transcription patterns of genes throughout the first three days of development in *M. galloprovincialis* and to obtain hints on the expected outcomes of mRNA-ISH experiments.

Raw reads were downloaded from the Sequence Read Archive (SRA) in NCBI (BioProject: PRJNA996031) and trimmed using Trimmomatic v0.39 (**Bolger et al., 2014**; `LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:65`). Read quality was checked using FastQC v0.12.1 (**Andrews et al., 2010**). Trimmed reads were mapped against the *M. galloprovincialis* annotated genome (GCA_900618805.1; **Gerdol et al., 2020**) using STAR v2.7.10b (**Dobin et al., 2013**) in `alignReads` mode with default parameters. The resulting gene count matrix was extracted with StringTie v2.2.1 (**Pertea et al., 2015, 2016**) in expression estimation mode followed by the python script `prepDE.py` (`-l 99`).

The resulting matrix was processed in R. Raw gene counts were normalized using the built-in function `vst` of the package DESeq2 (**Love et al., 2014**). The function `plotPCA` was then used to run a principal component analysis (PCA) on read mapping counts and visualize the corresponding results. Normalized gene counts were also used to plot expression values of target genes (i.e., *Vasa*, *Dmrt1L*, *SoxH* and *FoxL2*), as well as in maSigPro (**Conesa et al., 2006**) to run a differential gene expression analysis in a time course experiment.

The entire pipeline was automated through custom python and bash scripts, which are available in a private repository on GitHub.

4.2.2 Sample collection, MitoTracker staining and fixation

Adult Mediterranean mussels (*M. galloprovincialis*) were hand collected from various locations surrounding the AltaSea institute at the port of Los Angeles (CA, USA). Sampling took place during the late spawning season of the species in California, i.e., from October 2023 to early January 2024. Specimens were checked for species and sexual maturity before usage.

Selected mussels were thoroughly cleaned from epibionts and placed in ice for approximately 30-60 minutes, then transferred in filtered artificial sea water (FASW) at 16°C and acclimatized for 30 minutes. All the individuals were then placed in a common tank and spawning was induced by cyclical thermal shock, that is, by exposing mussels alternatively to FASW at 24-26°C and 14-16°C for 30-40 minutes. As soon as individual mussels started spawning, they were promptly removed from the common tank, carefully washed and then allowed to continue spawning in isolated containers of about 250 ml 16°C FASW.

Sperm from six males and oocytes from six females were separately mixed to increase the number of crosses. An hour after the spawning started, oocytes were filtered through a 75 over a 30 µm mesh and aged in 1 L of FASW for 40-60 minutes to let them assume a proper circular shape. Oocyte abundance was estimated under a stereo microscope, by counting the number of gametes in five aliquotes of 1 mL and then calculating the mean value. Sperm mitochondria were labeled with MitoTracker™ Red CMXRos (Thermo Fisher Scientific) at a working concentration of 500 nM for 30 minutes. MitoTracker is a vital and fixation-resistant mitochondrial dye and was used to be able to detect the sex of developing embryos (as early as the two-blastomere stage) according to the distribution pattern of sperm mitochondria (**Cao et al., 2004; Obata and Komaru, 2005**). From this step onward, samples were always kept in the dark.

Fertilization was performed by mixing oocytes and sperm at a ratio of 1:10. Fertilization success was checked after 20-30 minutes by the formation of polar bodies. The suspension was then carefully washed to remove excess sperm and brought to a concentration of 250 zygotes/mL. The resulting suspension was transferred into cell-culture flasks of 40 mL and embryos/larvae were reared at 16°C in the dark. Water was changed every 24 hours. After 48 hpf, larvae were fed with *Isochrysis galbana* at a final concentration of circa 100,000 cells/mL following **Helm et al., 2004**.

Embryos/larvae were sampled at 1, 2, 3 and 4 hpf, and then every 12 hours until 72 hpf,

Target	Amplifier	Fluorophore	No. of probe pairs
Vasa	B1	ALEXA-488	33
Dmrt1L	B2	ALEXA-647	18
SoxH	B3	ALEXA-546	22
FoxL2	B4	ALEXA-700	28

Table 4.1. List of genes targeted through HCR, with the corresponding amplifiers, fluorophores and number of generated probe pairs.

every time after checking for proper development and vitality. After concentration in a mesh of proper size, embryos/larvae were fixed in 3.2% paraformaldehyde (PFA) in 1× PBS at 4°C overnight under constant and gentle shaking. Fixed samples were washed 3 × 20 minutes in 1× PBS 0.1% Tween 20 (PBST) and then dehydrated 3 × 30 minutes in absolute methanol at room temperature (RT). Dehydrated samples were stored at -20°C until usage.

4.2.3 mRNA *in-situ* Hybridization Chain Reaction (HCR)

HCR probe design

Vasa, *Dmrt1L*, *SoxH*, and *FoxL2* spliced-transcript nucleotide sequences of *M. galloprovincialis* were obtained from previous analyses with OrthoFinder v2.5.5 (**Emms and Kelly, 2019**) and 30 annotated bivalve genomes (see **Chapter 3**). Accession numbers of spliced transcripts are 10B017427, 10B093608, 10B014180, and 10B094018, respectively. The `insitu_probe_generator` script from Ozpolat Lab (**Kuehn et al., 2022**) was used to generate pairs of probes specifically designed for third-generation HCR (**Choi et al., 2018**). The built-in BLASTN search against the annotated *M. galloprovincialis* transcriptome was employed to check for putative off-target bindings of probe pairs. B1-488, B2-647, B3-546, and B4-700 pairs of HCR amplifiers and fluorophores were chosen as in **Tab. 4.1**. Resulting probes were synthetized by Integrated DNA Techonologies (IDT™) in different oligo pools.

Fluorescent *in-situ* hybridization through hybridization chain reaction and microscope imaging

HCR mRNA-FISH in *M. galloprovincialis* embryos was performed following **Miglioli et al., 2024**. All the steps were carried out in the dark to prevent MitoTracker from fading. Probe hybridization buffer, probe wash buffer and amplification buffer were manufactured by Molecular Instruments, Inc.

Target	Dye	Excitation (nm)	Emission (nm)
dsDNA (nuclei)	DAPI	360	460
Sperm mitochondria	MitoTracker™ Red CMXRos	575	600
Vasa	ALEXA-488	499	520
Dmrt1L	ALEXA-647	653	670
SoxH	ALEXA-546	557	575
FoxL2	ALEXA-700	685	700

Table 4.2. List of dyes used for every target, together with the excitation and emission peaks as returned by the Las X software.

Dehydrated samples stored in methanol were washed 4 times per 5 minutes and 1 time per 10 minutes in a phosphate-buffered saline solution (PBS; 128 mM NaCl, 2 mM KCl, 8 mM $Na_2HPO_4 \cdot {}_2H_2O$, 2 mM KH_2PO_4) with 0.1% Tween 20 (PBST). Samples were then permeabilized for 30 minutes in a detergent solution (1.0% SDS, 0.5% Tween 20, 50 mM Tris-HCl, 1.0 mM ethylenediaminetetraacetic acid (EDTA), 150.0 mM NaCl) and washed again 2 times per 5 minutes in PBST. Samples were prepared for the HCR detection stage by incubation in probe hybridization buffer for 30 minutes at 37 °C. Detection stage was then performed with 4 nM of each probe set in hybridization solution overnight (>12 h) at 37 °C.

Excess probes was removed by washing 4 times per 20 minutes with probe wash buffer at 37 °C and 3 times per 5 minutes with 5× saline-sodium citrate Tween 20 buffer (SSCT; 5× SSC, 0.1% Tween 20) at room temperature. Samples were incubated for 30 minutes in amplification buffer at room temperature. Hairpins were heated at 95 °C for 90 seconds and then snap-cooled at room temperature for 30 minutes. The amplification step of HCR was performed with 6 pmol of each hairpin in amplification buffer overnight (>12 h) at room temperature.

Excess hairpins was removed by washing 2 times per 5 minutes, 2 times per 30 minutes, and 1 time per 5 minutes with SSCT. If not immediately mounted on slides, samples were stored in SSCT at +4 °C. Otherwise, samples were immersed in 50% and 75% glycerol for 30-60 minutes each, and then mounted with VECTASHIELD® PLUS Antifade Mounting Medium with DAPI (H-2000). Slides were imaged on a Stellaris 5 Confocal Package system with the software Las X (Leica Microsystems). Each dye was imaged sequentially in a separate channel, to enhance the yield and avoid any crosstalks. **Tab. 4.2** summarises the excitation and emission peaks for each dye. Images were then manipulated and post-produced using Fiji v2.14.0.

4.2.4 Immunolocalization of Vasa

Vasa immunolocalization in *M. galloprovincialis* embryos was performed following **Milani et al., 2011** with modifications. All the steps were carried out in the dark to prevent MitoTracker from fading.

Dehydrated samples stored in methanol were rinsed 3 times per 10 minutes and 1 time for 2 hours in Tris-buffered saline (TBS; 10 mM Tris-HCl, 155 mM NaCl), following an additional wash for 10 minutes with PBS. Samples were then digested for 6 minutes and 30 seconds with 0.01% pronase E (Merck) in PBS, and washed again 2 times for 5 minutes in PBS. Permeabilization was then performed in TBS-Triton (TBST) 0.1% for 5 minutes at RT and in TBST 1% overnight at 4°C.

After an additional rinse for 5 minutes in TBST 0.1%, non-specific protein-binding sites were blocked with a TBST 0.1% solution containing 3% bovine serum albumin (BSA). Samples were then incubated at 4°C for 32-48 hours with primary anti-VASA/VAS antibody (Abcam ab209710; polyclonal anti-Vasa developed in rabbit), diluted 1:100.

Excess primary antibody was rinsed from samples with 4 washes of 30 minutes in TBST 0.1%, while non-specific protein-binding sites were blocked again with an incubation of 1 hour in TBST 0.1% containing 3% BSA. Samples were then incubated at 4°C for 24-32 hours with secondary antibody HRP anti-rabbit in goat (Santa Cruz Biotechnology Inc.) diluited 1:400. Excess secondary antibody was rinsed with 4 washes of 30 minutes in TBST 0.1% and 1 wash of 1 hour in TBST 1%.

Samples were immersed in 50% and 75% glycerol for 30-60 minutes each, and then mounted with VECTASHIELD® PLUS Antifade Mounting Medium with DAPI (H-2000). Slides were imaged COMPLETECOMPLETECOMPLETECOMPLETE. Each dye was imaged sequentially in a separate channel, to enhance the yield and avoid any crosstalks. **Tab. 4.2** summarises the excitation and emission peaks for each dye. Images were then manipulated and post-produced using Fiji v2.14.0.

4.3 Results

In preparation.

4.4 Discussion

In preparation.

References

Abascal, F., Zardoya, R., & Telford, M. J. (2010). Translatorx: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic acids research*, *38*(suppl_2), W7–W13.

Abbott, J. K. (2011). Intra-locus sexual conflict and sexually antagonistic genetic variation in hermaphroditic animals. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1703), 161–169.

Alexa, A., & Rahnenführer, J. (2009). Gene set enrichment analysis with topgo. *Bioconductor Improv*, *27*, 1–26.

Altenhoff, A. M., Warwick Vesztrocy, A., Bernard, C., Train, C.-M., Nicheperovich, A., Prieto Baños, S., Julca, I., Moi, D., Nevers, Y., Majidian, S., et al. (2024). Oma orthology in 2024: Improved prokaryote coverage, ancestral and extant go enrichment, a revamped synteny viewer and more in the oma ecosystem. *Nucleic Acids Research*, *52*(D1), D513–D521.

Andrews, S., et al. (2010). Fastqc: A quality control tool for high throughput sequence data.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., et al. (2014). Sex determination: Why so many ways of doing it? *PLoS biology*, *12*(7), e1001899.

Bai, C.-M., Xin, L.-S., Rosani, U., Wu, B., Wang, Q.-C., Duan, X.-K., Liu, Z.-H., & Wang, C.-M. (2019). Chromosomal-level assembly of the blood clam, *Scapharca* (*Anadara*) *broughtonii*, using long sequence reads and hi-c. *GigaScience*, *8*(7), giz067.

Baldwin-Brown, J. G., Weeks, S. C., & Long, A. D. (2018). A new standard for crustacean genomes: The highly contiguous, annotated genome assembly of the clam shrimp *Eulimnadia texana* reveals hox gene order and identifies the sex chromosome. *Genome biology and evolution*, *10*(1), 143–156.

Baral, S., Arumugam, G., Deshmukh, R., & Kunte, K. (2019). Genetic architecture and sex-specific selection govern modular, male-biased evolution of *doublesex*. *Science advances*, *5*(5), eaau3753.

Beukeboom, L. W., & Perrin, N. (2014). *The evolution of sex determination*. Oxford University Press.

Bewick, A. J., Anderson, D. W., & Evans, B. J. (2011). Evolution of the closely related, sex-related genes *DM-W* and *DMRT1* in african clawed frogs (*Xenopus*). *Evolution*, *65*(3), 698–712.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bredemeyer, K. R., Hillier, L., Harris, A. J., Hughes, G. M., Foley, N. M., Lawless, C., Carroll, R. A., Storer, J. M., Batzer, M. A., Rice, E. S., et al. (2023). Single-haplotype comparative genomics provides insights into lineage-specific structural variation during cat evolution. *Nature genetics*, *55*(11), 1953–1963.

Breton, S., Capt, C., Guerra, D., & Stewart, D. (2018). Sex-determining mechanisms in bivalves. In J. L. Leonard (Ed.), *Transitions between sexual systems: Understanding the mechanisms of, and pathways between, dioecy, hermaphroditism and other sexual systems* (pp. 165–192). Springer International Publishing.

Breton, S., Stewart, D. T., Brémaud, J., Havird, J. C., Smith, C. H., & Hoeh, W. R. (2022). Did doubly uniparental inheritance (dui) of mtdna originate as a cytoplasmic male sterility (cms) system? *BioEssays*, *44*(4), 2100283.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, *12*(1), 59–60.

Calcino, A. D., de Oliveira, A. L., Simakov, O., Schwaha, T., Zieger, E., Wollesen, T., & Wanninger, A. (2019). The quagga mussel genome and the evolution of freshwater tolerance. *DNA Research*, *26*(5), 411–422.

Cao, L., Kenchington, E., & Zouros, E. (2004). Differential segregation patterns of sperm mitochondria in embryos of the blue mussel (*Mytilus edulis*). *Genetics*, *166*(2), 883–894.

Capel, B. (2017). Vertebrate sex determination: Evolutionary plasticity of a fundamental switch. *Nature Reviews Genetics*, *18*(11), 675–689.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). Trimal: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973.

Capt, C., Bouvet, K., Guerra, D., Robicheau, B. M., Stewart, D. T., Pante, E., & Breton, S. (2020). Unorthodox features in two venerid bivalves with doubly uniparental inheritance of mitochondria. *Scientific reports*, *10*(1), 1087.

Capt, C., Renaut, S., Ghiselli, F., Milani, L., Johnson, N. A., Sietman, B. E., Stewart, D. T., & Breton, S. (2018). Deciphering the link between doubly uniparental inheritance of mtdna and sex determination in bivalves: Clues from comparative transcriptomics. *Genome biology and evolution*, *10*(2), 577–590.

Chen, H., Xiao, G., Chai, X., Lin, X., Fang, J., & Teng, S. (2017). Transcriptome analysis of sex-related genes in the blood clam *Tegillarca granosa*. *PLoS One*, *12*(9), e0184584.

Choi, H. M., Schwarzkopf, M., Fornace, M. E., Acharya, A., Artavanis, G., Stegmaier, J., Cunha, A., & Pierce, N. A. (2018). Third-generation in situ hybridization chain reaction: Multiplexed, quantitative, sensitive, versatile, robust. *Development*, *145*(12), dev165753.

Collin, R. (2013). Phylogenetic patterns and phenotypic plasticity of molluscan sexual systems. *Integrative and Comparative Biology*, *53*(4), 723–735.

Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). Masigpro: A method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, *22*(9), 1096–1102.

Corrochano-Fraile, A., Davie, A., Carboni, S., & Bekaert, M. (2022). Evidence of multiple genome duplication events in *Mytilus* evolution. *BMC genomics*, *23*(1), 340.

Cutter, A. D., & Ward, S. (2005). Sexual and temporal dynamics of molecular evolution in *C. elegans* development. *Molecular Biology and Evolution*, *22*(1), 178–188.

Deng, T., Pang, C., Lu, X., Zhu, P., Duan, A., Tan, Z., Huang, J., Li, H., Chen, M., & Liang, X. (2016). *De novo* transcriptome assembly of the chinese swamp buffalo by rna sequencing and ssr marker discovery. *PLoS One*, *11*(1), e0147132.

Dermauw, W., Van Leeuwen, T., & Feyereisen, R. (2020). Diversity and evolution of the p450 family in arthropods. *Insect biochemistry and molecular biology*, *127*, 103490.

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*(1), 15–21.

Eads, B. D., Colbourne, J. K., Bohuski, E., & Andrews, J. (2007). Profiling sex-biased gene expression during parthenogenetic reproduction in *Daphnia pulex*. *BMC genomics*, *8*, 1–14.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS computational biology*, *7*(10), e1002195.

Elbers, J. P., Rogers, M. F., Perelman, P. L., Proskuryakova, A. A., Serdyukova, N. A., Johnson, W. E., Horin, P., Corander, J., Murphy, D., & Burger, P. A. (2019). Improving illumina assemblies with hi-c and long reads: An example with the north african dromedary. *Molecular Ecology Resources*, *19*(4), 1015–1026.

Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*, 1–14.

Evensen, K. G., Robinson, W. E., Krick, K., Murray, H. M., & Poynton, H. C. (2022). Comparative phylotranscriptomics reveals putative sex differentiating genes across eight diverse bivalve species. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *41*, 100952.

Fan, H., Wu, Q., Wei, F., Yang, F., Ng, B. L., & Hu, Y. (2019). Chromosome-level genome assembly for giant panda provides novel insights into carnivora chromosome evolution. *Genome biology*, *20*, 1–12.

Foote, A. D., Liu, Y., Thomas, G. W., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C. E., Hunter, M. E., Joshi, V., et al. (2015). Convergent evolution of the genomes of marine mammals. *Nature genetics*, *47*(3), 272–275.

Fouret, J., Brunet, F. G., Binet, M., Aurine, N., Enchéry, F., Croze, S., Guinier, M., Goumaidi, A., Preininger, D., Volff, J.-N., et al. (2020). Sequencing the genome of indian flying fox, natural reservoir of nipah virus, using hybrid assembly and conservative secondary scaffolding. *Frontiers in Microbiology*, *11*, 1807.

Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-Ortiz, M. A., Murgarella, M., Greco, S., et al. (2020). Massive gene presence-absence variation shapes an open pan-genome in the mediterranean mussel. *Genome biology*, *21*, 1–21.

Ghiselli, F., Iannello, M., Piccinini, G., & Milani, L. (2021). Bivalve molluscs as model systems for studying mitochondrial biology. *Integrative and Comparative Biology*, *61*(5), 1699–1714.

Ghiselli, F., Iannello, M., Puccio, G., Chang, P. L., Plazzi, F., Nuzhdin, S. V., & Passamonti, M. (2018). Comparative transcriptomics in two bivalve species offers different perspectives on the evolution of sex-biased genes. *Genome Biology and Evolution*, *10*(6), 1389–1402.

Ghiselli, F., Milani, L., Guerra, D., Chang, P. L., Breton, S., Nuzhdin, S. V., & Passamonti, M. (2013). Structure, transcription, and variability of metazoan mitochondrial genome: Perspectives from an unusual mitochondrial inheritance system. *Genome biology and evolution*, *5*(8), 1535–1554.

Gomes-dos-Santos, A., Lopes-Lima, M., Machado, A. M., Marcos Ramos, A., Usié, A., Bolotov, I. N., Vikhrev, I. V., Breton, S., Castro, L. F. C., da Fonseca, R. R., et al. (2021). The crown pearl: A draft genome assembly of the european freshwater pearl mussel *Margaritifera margaritifera* (linnaeus, 1758). *DNA research*, *28*(2), dsab002.

Gómez-Chiarri, M., Warren, W. C., Guo, X., & Proestou, D. (2015). Developing tools for the study of molluscan immunity: The sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish & shellfish immunology*, *46*(1), 2–4.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology*, *29*(7), 644.

Grath, S., & Parsch, J. (2016). Sex-biased gene expression. *Annual review of genetics*, *50*, 29–44.

Grau-Bové, X., & Sebé-Pedrós, A. (2021). Orthology clusters from gene trees with possvm. *Molecular Biology and Evolution*, *38*(11), 5204–5208.

Gusman, A., Lecomte, S., Stewart, D. T., Passamonti, M., & Breton, S. (2016). Pursuing the quest for better understanding the taxonomic distribution of the system of doubly uniparental inheritance of mtdna. *PeerJ*, *4*, e2760.

Han, F., Wang, Z., Wu, F., Liu, Z., Huang, B., & Wang, D. (2010). Characterization, phylogeny, alternative splicing and expression of sox30 gene. *BMC Molecular Biology*, *11*, 1–11.

Han, W., Liu, L., Wang, J., Wei, H., Li, Y., Zhang, L., Guo, Z., Li, Y., Liu, T., Zeng, Q., et al. (2022). Ancient homomorphy of molluscan sex chromosomes sustained by reversible sex-biased genes and sex determiner translocation. *Nature Ecology & Evolution*, 1–16.

Heenan, P., Zondag, L., & Wilson, M. J. (2016). Evolution of the sox gene family within the chordate phylum. *Gene*, *575*(2), 385–392.

Helm, M. M., Bourne, N., & Lovatelli, A. (2004). *Hatchery culture of bivalves: A practical manual*.

Iannello, M., Forni, G., Piccinini, G., Xu, R., Martelossi, J., Ghiselli, F., & Milani, L. (2023). Signatures of extreme longevity: A perspective from bivalve molecular evolution. *Genome Biology and Evolution*, *15*(11), evad159.

Inoue, K., Yoshioka, Y., Tanaka, H., Kinjo, A., Sassa, M., Ueda, I., Shinzato, C., Toyoda, A., & Itoh, T. (2021). Genomics and transcriptomics of the green mussel explain the durability of its byssus. *Scientific Reports*, *11*(1), 5992.

Ip, J. C.-H., Xu, T., Sun, J., Li, R., Chen, C., Lan, Y., Han, Z., Zhang, H., Wei, J., Wang, H., et al. (2021). Host–endosymbiont genome integration in a deep-sea chemosymbiotic clam. *Molecular Biology and Evolution*, *38*(2), 502–518.

Jackson, B. C., Carpenter, C., Nebert, D. W., & Vasiliou, V. (2010). Update of human and mouse forkhead box (fox) gene families. *Human genomics*, *4*, 1–8.

Jombart, T., & Dray, S. (2010). Adephylo: Exploratory analyses for the phylogenetic comparative method. *Bioinformatics*, *26*(15), 1–21.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). Modelfinder: Fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), 587–589.

Kenny, N. J., McCarthy, S. A., Dudchenko, O., James, K., Betteridge, E., Corton, C., Dolucan, J., Mead, D., Oliver, K., Omer, A. D., et al. (2020). The gene-rich genome of the scallop *Pecten maximus*. *Gigascience*, *9*(5), giaa037.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., & Paabo, S. (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, *309*(5742), 1850–1854.

Kopp, A. (2012). Dmrt genes in the development and evolution of sexual dimorphism. *Trends in Genetics*, *28*(4), 175–184.

Kousathanas, A., Halligan, D. L., & Keightley, P. D. (2014). Faster-x adaptive protein evolution in house mice. *Genetics*, *196*(4), 1131–1143.

Kuehn, E., Clausen, D. S., Null, R. W., Metzger, B. M., Willis, A. D., & Özpolat, B. D. (2022). Segment number threshold determines juvenile onset of germline cluster expansion in platynereis dumerilii. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *338*(4), 225–240.

Larroux, C., Luke, G. N., Koopman, P., Rokhsar, D. S., Shimeld, S. M., & Degnan, B. M. (2008). Genesis and expansion of metazoan transcription factor gene classes. *Molecular biology and evolution*, *25*(5), 980–996.

Lemoine, F., & Gascuel, O. (2021). Gotree/goalign: Toolkit and go api to facilitate the development of phylogenetic workflows. *NAR Genomics and Bioinformatics*, *3*(3), lqab075.

Li, A., Dai, H., Guo, X., Zhang, Z., Zhang, K., Wang, C., Wang, X., Wang, W., Chen, H., Li, X., et al. (2021). Genome of the estuarine oyster provides insights into climate impact and adaptive plasticity. *Communications Biology*, *4*(1), 1287.

Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics*, *39*(1), btad014.

Li, R., Zhang, L., Li, W., Zhang, Y., Li, Y., Zhang, M., Zhao, L., Hu, X., Wang, S., & Bao, Z. (2018). *FOXL2* and *DMRT1L* are yin and yang genes for determining timing of sex differentiation in the bivalve mollusk *Patinopecten yessoensis*. *Frontiers in Physiology*, *9*, 1166.

Liang, S., Liu, D., Li, X., Wei, M., Yu, X., Li, Q., Ma, H., Zhang, Z., & Qin, Z. (2019). *SOX2* participates in spermatogenesis of zhikong scallop *Chlamys farreri*. *Scientific Reports*, *9*(1), 76.

Liu, T., Zhang, Y., Nie, H., Sun, J., & Yan, X. (2024). Characterization and expression patterns of the fox gene family under heat and cold stress in manila clam *Ruditapes philippinarum* based on genome-wide identification. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 101313.

Liu, X., Li, C., Chen, M., Liu, B., Yan, X., Ning, J., Ma, B., Liu, G., Zhong, Z., Jia, Y., et al. (2020). Draft genomes of two atlantic bay scallop subspecies *Argopecten irradians irradians* and *A. i. concentricus*. *Scientific Data*, *7*(1), 99.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, *15*(12), 1–21.

Lozano-Fernandez, J. (2022). A practical guide to design and assess a phylogenomic study. *Genome Biology and Evolution*, *14*(9), evac129.

Mank, J. E., Axelsson, E., & Ellegren, H. (2007). Fast-x on the z: Rapid evolution of sex-linked genes in birds. *Genome Research*, *17*(5), 618–624.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). Busco update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution*, *38*(10), 4647–4654.

Marshall Graves, J. A., & Peichel, C. L. (2010). Are homologies in vertebrate sex determination due to shared ancestry or to limited options? *Genome biology*, *11*, 1–12.

Mawaribuchi, S., Ito, Y., & Ito, M. (2019). Independent evolution for sex determination and differentiation in the *DMRT* family in animals. *Biology Open*, *8*(8), bio041962.

Mazet, F., Yu, J.-K., Liberles, D. A., Holland, L. Z., & Shimeld, S. M. (2003). Phylogenetic relationships of the fox (forkhead) gene family in the bilateria. *Gene*, *316*, 79–89.

McCartney, M. A., Auch, B., Kono, T., Mallez, S., Zhang, Y., Obille, A., Becker, A., Abrahante, J. E., Garbe, J., Badalamenti, J. P., et al. (2022). The genome of the zebra mussel, *Dreissena polymorpha*: A resource for comparative genomics, invasion genetics, and biocontrol. *G3*, *12*(2), jkab423.

Mead, D., Fingland, K., Cripps, R., Miguez, R. P., Smith, M., Corton, C., Oliver, K., Skelton, J., Betteridge, E., Doulcan, J., et al. (2020). The genome sequence of the eastern grey squirrel, *Sciurus carolinensis* gmelin, 1788. *Wellcome Open Research*, *5*.

Meisel, R. P., & Connallon, T. (2013). The faster-x effect: Integrating theory and data. *Trends in genetics*, *29*(9), 537–544.

Miglioli, A., Tredez, M., Boosten, M., Sant, C., Carvalho, J. E., Dru, P., Canesi, L., Schubert, M., & Dumollard, R. (2024). The mediterranean mussel mytilus galloprovincialis: A novel model for developmental studies in mollusks. *Development*, *151*(4), dev202256.

Milani, L., & Ghiselli, F. (2020). Faraway, so close. the comparative method and the potential of non-model animals in mitochondrial research. *Philosophical Transactions of the Royal Society B*, *375*(1790), 20190186.

Milani, L., Ghiselli, F., Maurizii, M. G., Nuzhdin, S. V., & Passamonti, M. (2014). Paternally transmitted mitochondria express a new gene of potential viral origin. *Genome biology and evolution*, *6*(2), 391–405.

Milani, L., Ghiselli, F., Maurizii, M. G., & Passamonti, M. (2011). Doubly uniparental inheritance of mitochondria as a model system for studying germ line formation. *PLoS One*, *6*(11), e28194.

Milani, L., Ghiselli, F., Nuzhdin, S. V., & Passamonti, M. (2013). Nuclear genes with sex bias in *Ruditapes philippinarum* (bivalvia, veneridae): Mitochondrial inheritance and sex determination in dui species. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *320*(7), 442–454.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), 1530–1534.

Moreno, J. A., Dudchenko, O., Feigin, C. Y., Mereby, S. A., Chen, Z., Ramos, R., Almet, A. A., Sen, H., Brack, B. J., Johnson, M. R., et al. (2024). *Emx2* underlies the development and evolution of marsupial gliding membranes. *Nature*, 1–9.

Murata, Y., Nikaido, M., Sasaki, T., Cao, Y., Fukumoto, Y., Hasegawa, M., & Okada, N. (2003). Afrotherian phylogeny as inferred from complete mitochondrial genomes. *Molecular phylogenetics and evolution*, *28*(2), 253–260.

Nakagawa, S., Gisselbrecht, S. S., Rogers, J. M., Hartl, D. L., & Bulyk, M. L. (2013). Dna-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences*, *110*(30), 12349–12354.

Natsidis, P., Kapli, P., Schiffer, P. H., & Telford, M. J. (2021). Systematic errors in orthology inference and their effects on evolutionary analyses. *Iscience*, *24*(2).

Nicolini, F., Martelossi, J., Forni, G., Savojardo, C., Mantovani, B., & Luchetti, A. (2023). Comparative genomics of *Hox* and *ParaHox* genes among major lineages of branchiopoda with emphasis on tadpole shrimps. *Frontiers in Ecology and Evolution*, *11*, 1046960.

Obata, M., & Komaru, A. (2005). Specific location of sperm mitochondria in mussel mytilus galloprovincialis zygotes stained by mitotracker. *Development, growth & differentiation*, *47*(4), 255–263.

of Sex Consortium, T., et al. (2014). Tree of sex: A database of sexual systems. *Scientific Data*, *1*.

Orkin, J. D., Montague, M. J., Tejada-Martinez, D., De Manuel, M., Del Campo, J., Cheves Hernandez, S., Di Fiore, A., Fontsere, C., Hodgson, J. A., Janiak, M. C., et al. (2021). The genomics of ecological flexibility, large brains, and long lives in capuchin monkeys revealed with fecalfacs. *Proceedings of the National Academy of Sciences*, *118*(7), e2010632118.

Panara, V., Budd, G. E., & Janssen, R. (2019). Phylogenetic analysis and embryonic expression of panarthropod dmrt genes. *Frontiers in zoology*, *16*, 1–18.

Papa, F., Windbichler, N., Waterhouse, R. M., Cagnetti, A., D'Amato, R., Persampieri, T., Lawniczak, M. K., Nolan, T., & Papathanos, P. A. (2017). Rapid evolution of female-biased genes among four species of *Anopheles* malaria mosquitoes. *Genome research*, *27*(9), 1536–1548.

Parsch, J., & Ellegren, H. (2013). The evolutionary causes and consequences of sex-biased gene expression. *Nature Reviews Genetics*, *14*(2), 83–87.

Peñaloza, C., Gutierrez, A. P., Eöry, L., Wang, S., Guo, X., Archibald, A. L., Bean, T. P., & Houston, R. D. (2021). A chromosome-level genome assembly for the pacific oyster *Crassostrea gigas*. *GigaScience*, *10*(3), giab020.

Perez-Garcia, C., Moran, P., & Pasantes, J. J. (2011). Cytogenetic characterization of the invasive mussel species *Xenostrobus securis* lmk. (bivalvia: Mytilidae). *Genome*, *54*(09), 771–778.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature protocols*, *11*(9), 1650–1667.

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*(3), 290–295.

Phochanukul, N., & Russell, S. (2010). No backbone but lots of sox: Invertebrate sox genes. *The international journal of biochemistry & cell biology*, *42*(3), 453–464.

Piccinini, G., Iannello, M., Puccio, G., Plazzi, F., Havird, J. C., & Ghiselli, F. (2021). Mitonuclear coevolution, but not nuclear compensation, drives evolution of oxphos complexes in bivalves. *Molecular Biology and Evolution*, *38*(6), 2597–2614.

Powell, D., Subramanian, S., Suwansa-Ard, S., Zhao, M., O'Connor, W., Raftos, D., & Elizur, A. (2018). The genome of the oyster saccostrea offers insight into the environmental resilience of bivalves. *DNA Research*, *25*(6), 655–665.

Purandare, S. R., Bickel, R. D., Jaquiery, J., Rispe, C., & Brisson, J. A. (2014). Accelerated evolution of morph-biased genes in pea aphids. *Molecular biology and evolution*, *31*(8), 2073–2083.

Ran, Z., Li, Z., Yan, X., Liao, K., Kong, F., Zhang, L., Cao, J., Zhou, C., Zhu, P., He, S., et al. (2019). Chromosome-level genome assembly of the razor clam *Sinonovacula constricta* (lamarck, 1818). *Molecular ecology resources*, *19*(6), 1647–1658.

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). Mafft-dash: Integrated protein sequence and structural alignment. *Nucleic acids research*, *47*(W1), W5–W10.

Sarkar, A., & Hochedlinger, K. (2013). The *sox* family of transcription factors: Versatile regulators of stem and progenitor cell fate. *Cell stem cell*, *12*(1), 15–30.

Schliep, K. P. (2011). Phangorn: Phylogenetic analysis in r. *Bioinformatics*, *27*(4), 592–593.

Schomburg, C., Janssen, R., & Prpic, N.-M. (2022). Phylogenetic analysis of forkhead transcription factors in the panarthropoda. *Development genes and evolution*, *232*(1), 39–48.

Seudre, O., Martín-Zamora, F. M., Rapisarda, V., Luqman, I., Carrillo-Baltodano, A. M., & Martín-Durán, J. M. (2022). The fox gene repertoire in the annelid *Owenia fusiformis* reveals multiple expansions of the *foxQ2* class in spiralia. *Genome Biology and Evolution*, *14*(10), evac139.

Shi, J., Hong, Y., Sheng, J., Peng, K., & Wang, J. (2015). De novo transcriptome sequencing to identify the sex-determination genes in *Hyriopsis schlegelii*. *Bioscience, Biotechnology, and Biochemistry*, *79*(8), 1257–1265.

Shi, Y., Liu, W., & He, M. (2018). Proteome and transcriptome analysis of ovary, intersex gonads, and testis reveals potential key sex reversal/differentiation genes and mechanism in scallop *Chlamys nobilis*. *Marine Biotechnology*, *20*, 220–245.

Shimeld, S. M., Boyle, M. J., Brunet, T., Luke, G. N., & Seaver, E. C. (2010). Clustered fox genes in lophotrochozoans and the evolution of the bilaterian fox gene cluster. *Developmental biology*, *340*(2), 234–248.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, *7*(1), 539.

Smith, C. H. (2021). A high-quality reference genome for a parasitic bivalve with doubly uniparental inheritance (bivalvia: Unionida). *Genome Biology and Evolution*, *13*(3), evab029.

Song, H., Guo, X., Sun, L., Wang, Q., Han, F., Wang, H., Wray, G. A., Davidson, P., Wang, Q., Hu, Z., et al. (2021). The hard clam genome reveals massive expansion and diversification of inhibitors of apoptosis in bivalvia. *BMC biology*, *19*, 1–20.

Stammnitz, M. R., Gori, K., Kwon, Y. M., Harry, E., Martin, F. J., Billis, K., Cheng, Y., Baez-Ortega, A., Chow, W., Comte, S., et al. (2023). The evolution of two transmissible cancers in tasmanian devils. *Science*, *380*(6642), 283–293.

Stothard, P., & Pilgrim, D. (2003). Sex-determination gene and pathway evolution in nematodes. *Bioessays*, *25*(3), 221–231.

Sun, D., Yu, H., & Li, Q. (2022). Examination of the roles of *Foxl2* and *Dmrt1* in sex differentiation and gonadal development of oysters by using rna interference. *Aquaculture*, *548*, 737732.

Tu, Q., Brown, C. T., Davidson, E. H., & Oliveri, P. (2006). Sea urchin forkhead gene family: Phylogeny and embryonic expression. *Developmental biology*, *300*(1), 49–62.

Uller, T., & Helanterä, H. (2011). From the origin of sex-determining factors to the evolution of sex-determining systems. *The Quarterly review of biology*, *86*(3), 163–180.

Verhulst, E. C., van de Zande, L., & Beukeboom, L. W. (2010). Insect sex determination: It all evolves around transformer. *Current opinion in genetics & development*, *20*(4), 376–383.

Vicoso, B., & Charlesworth, B. (2006). Evolution on the x chromosome: Unusual patterns and processes. *Nature Reviews Genetics*, *7*(8), 645–653.

Vilstrup, J. T., Seguin-Orlando, A., Stiller, M., Ginolhac, A., Raghavan, M., Nielsen, S. C., Weinstock, J., Froese, D., Vasiliev, S. K., Ovodov, N. D., et al. (2013). Mitochondrial phylogenomics of modern and ancient equids. *PloS one*, *8*(2), e55950.

Vizueta Moraga, J., Sánchez-Gracia, A., & Rozas Liras, J. A. (2020). Bitacora: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Molecular Ecology Resources, 2020, vol. 20, num. 5, p. 1445-1452.*

Wang, C., Wallerman, O., Arendt, M.-L., Sundström, E., Karlsson, Å., Nordin, J., Mäkeläinen, S., Pielberg, G. R., Hanson, J., Ohlsson, Å., et al. (2021). A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Communications biology*, *4*(1), 185.

Wang, G., Dong, S., Guo, P., Cui, X., Duan, S., & Li, J. (2020). Identification of *Foxl2* in freshwater mussel *Hyriopsis cumingii* and its involvement in sex differentiation. *Gene*, *754*, 144853.

Wang, J., & Nie, H. (2024). Genome-wide identification and expression analysis of sox gene family in the manila clam (ruditapes philippinarum). *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *50*, 101244.

Wang, M., Xia, J., Jawad, M., Wei, W., Gui, L., Liang, X., Yang, J.-L., & Li, M. (2022). Transcriptome sequencing analysis of sex-related genes and mirnas in the gonads of mytilus coruscus. *Frontiers in Marine Science*, *9*, 1013857.

Wang, Q., Cao, T., & Wang, C. (2023). Genome-wide identification and expression analysis of dmrt genes in bivalves. *BMC genomics*, *24*(1), 457.

Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., Guo, X., Huan, P., Dong, B., Zhang, L., et al. (2017). Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature ecology & evolution*, *1*(5), 0120.

Wang, X., Werren, J. H., & Clark, A. G. (2015). Genetic and epigenetic architecture of sex-biased expression in the jewel wasps *Nasonia vitripennis* and *girault*i. *Proceedings of the National Academy of Sciences*, *112*(27), E3545–E3554.

Wei, M., Ge, H., Shao, C., Yan, X., Nie, H., Duan, H., Liao, X., Zhang, M., Chen, Y., Zhang, D., et al. (2020). Chromosome-level clam genome helps elucidate the molecular basis of adaptation to a buried lifestyle. *IScience*, *23*(6).

Wexler, J. R., Plachetzki, D. C., & Kopp, A. (2014). Pan-metazoan phylogeny of the dmrt gene family: A framework for functional studies. *Development Genes and Evolution*, *224*, 175–181.

Willson, J., Roddur, M. S., Liu, B., Zaharias, P., & Warnow, T. (2022). Disco: Species tree inference using multicopy gene family tree decomposition. *Systematic biology*, *71*(3), 610–629.

Wu, S., Zhang, Y., Li, Y., Wei, H., Guo, Z., Wang, S., Zhang, L., & Bao, Z. (2020). Identification and expression profiles of fox transcription factors in the yesso scallop (patinopecten yessoensis). *Gene*, *733*, 144387.

Xiong, Y., Brandley, M. C., Xu, S., Zhou, K., & Yang, G. (2009). Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. *BMC Evolutionary Biology*, *9*, 1–13.

Xu, R., Iannello, M., Havird, J. C., Milani, L., & Ghiselli, F. (2022). Lack of transcriptional coordination between mitochondrial and nuclear oxidative phosphorylation genes in the presence of two divergent mitochondrial genomes. *Zoological Research*, *43*(1), 111.

Xu, R., Martelossi, J., Smits, M., Iannello, M., Peruzza, L., Babbucci, M., Milan, M., Dunham, J. P., Breton, S., Milani, L., et al. (2022). Multi-tissue rna-seq analysis and long-read-based genome assembly reveal complex sex-specific gene regulation and molecular evolution in the manila clam. *Genome Biology and Evolution*, *14*(12), evac171.

Yang, J.-L., Feng, D.-D., Liu, J., Xu, J.-K., Chen, K., Li, Y.-F., Zhu, Y.-T., Liang, X., & Lu, Y. (2021). Chromosome-level genome assembly of the hard-shelled mussel *Mytilus coruscus*, a widely distributed species from the temperate areas of east asia. *Gigascience*, *10*(4), giab024.

Yang, M., Xu, F., Liu, J., Que, H., Li, L., & Zhang, G. (2014). Phylogeny of forkhead genes in three spiralians and their expression in pacific oyster *Crassostrea gigas*. *Chinese journal of oceanology and limnology*, *32*(6), 1207–1223.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36.

Yu, J., Zhang, L., Li, Y., Li, R., Zhang, M., Li, W., Xie, X., Wang, S., Hu, X., & Bao, Z. (2017). Genome-wide identification and expression profiling of the sox gene family in a bivalve mollusc *Patinopecten yessoensis*. *Gene*, *627*, 530–537.

Yu, J.-K., Mazet, F., Chen, Y.-T., Huang, S.-W., Jung, K.-C., & Shimeld, S. M. (2008). The fox genes of *Branchiostoma floridae*. *Development genes and evolution*, *218*, 629–638.

Yue, C., Li, Q., & Yu, H. (2021). Variance in expression and localization of sex-related genes *CgDsx*, *CgBHMG1* and *CgFoxl2* during diploid and triploid pacific oyster *Crassostrea gigas* gonad differentiation. *Gene*, *790*, 145692.

Zeng, Y., Zheng, H., He, C., Zhang, C., Zhang, H., & Zheng, H. (2024). Genome-wide identification and expression analysis of dmrt gene family and their role in gonad development of pacific oyster (crassostrea gigas). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *269*, 110904.

Zhang, N., Xu, F., & Guo, X. (2014). Genomic analysis of the pacific oyster (*Crassostrea gigas*) reveals possible conservation of vertebrate sex determination in a mollusc. *G3: Genes, Genomes, Genetics*, *4*(11), 2207–2217.

Appendix

The appendix includes the titles and abstracts of the papers published during my PhD that are not part of this thesis.

Taxonomic revision of the Australian stick insect genus *Candovia* (Phasmida: Necrosciinae): insight from molecular systematics and species-delimitation approaches.

Giobbe Forni[1,2], Alex Cussigh[1,2], Paul D. Brock[3], Braxton R. Jones[4], Filippo Nicolini[1], Jacopo Martelossi[1], Andrea Luchetti[1], Barbara Mantovani[1]

[1]*Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy.*

[2]*Department of Agricultural and Environmental Sciences, University of Milan, Milano, Italy.*

[3]*The Natural History Museum, Cromwell Road, London, UK.*

[4]*School of Life and Environmental Sciences, The University of Sydney, Sydney NSW 2006, Australia.*

Published in: 2023, *Zoological Journal of the Linnean Society*, 197:189–210. 10.1093/zoolinnean/zlac074

Abstract. The Phasmida genus *Candovia* comprises nine traditionally recognized species, all endemic to Australia. In this study, *Candovia* diversity is explored through molecular species-delimitation analyses using the COI_{Fol} gene fragment and phylogenetic inferences leveraging seven additional mitochondrial and nuclear loci. Molecular results were integrated with morphological observations, leading us to confirm the already described species and to the delineation of several new taxa and of the new genus *Paracandovia*. New *Candovia* species from various parts of Queensland and New South Wales are described and illustrated (*C. alata* sp. nov., *C. byfieldensis* sp. nov., *C. dalgleishae* sp. nov., *C. eungellensis* sp. nov., *C. karasi* sp. nov., *C. koensi* sp. nov. and *C. wollumbinensis* sp. nov.). New combinations are proposed and species removed from synonymy with the erection of the new genus *Paracandovia* (*P. cercata* stat. rev., comb. nov., *P. longipes* stat. rev., comb. nov., *P. pallida* comb. nov., *P. peridromes* comb. nov., *P. tenera* stat. rev., comb. nov.). Phylogenetic analyses suggest that the egg capitulum may have independently evolved multiple times throughout the evolutionary history of these insects. Furthermore, two newly described species represent the first taxa with fully

developed wings in this previously considered apterous clade.

Comparative genomics of *Hox* and *ParaHox* genes among major lineages of Branchiopoda with emphasis on tadpole shrimps.

Filippo Nicolini[1,2], Jacopo Martelossi[1], Giobbe Forni[3], Castrense Savojardo[4], Barbara Mantovani[1], Andrea Luchetti[1]

[1]*Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy.*

[2]*Fano Marine Center, Fano (PU), Italy.*

[3]*Department of Agricultural and Environmental Sciences, University of Milan, Milan, Italy.*

[4]*Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy.*

Published in: 2023, *Frontiers in Ecology and Evolution*, 11:1046960. 10.3389/fevo.2023.1046960

Abstract. *Hox* and *ParaHox* genes (HPHGs) are key developmental genes that pattern regional identity along the anterior–posterior body axis of most animals. Here, we identified HPHGs in tadpole shrimps (Pancrustacea, Branchiopoda, Notostraca), an iconic example of the so-called "living fossils" and performed a comparative genomics analysis of HPHGs and the *Hox* cluster among major branchiopod lineages. Notostraca possess the entire *Hox* complement, and the *Hox* cluster seems to be split into two different subclusters, although we were not able to support this finding with chromosome-level assemblies. However, the genomic structure of *Hox* genes in Notostraca appears more derived than that of *Daphnia* spp., which instead retains the plesiomorphic condition of a single compact cluster. Spinicaudata and *Artemia franciscana* show instead a *Hox* cluster subdivided across two or more genomic scaffolds with some orthologs either duplicated or missing. Yet, branchiopod HPHGs are similar among the various clades in terms of both intron length and number, as well as in their pattern of molecular evolution. Sequence substitution rates are in fact generally similar for most of the branchiopod *Hox* genes and the few differences we found cannot be traced back to natural selection, as they are not associated with any signals of diversifying selection or substantial switches in selective modes. Altogether, these findings do not support a significant stasis in the Notostraca *Hox* cluster and further confirm how morphological evolution is not tightly associated with genome dynamics.

Multiple and diversified transposon lineages contribute to early and recent bivalve genome evolution.

Jacopo Martelossi[1], Filippo Nicolini[1,2], Simone Subacchi[1], Daniela Pasquale[1], Fabrizio Ghiselli[1], Andrea Luchetti[1]

[1]*Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy.*

[2]*Fano Marine Center, Fano (PU), Italy.*

Published in: 2023, *BMC Biology*, 21:145.
10.1186/s12915-023-01632-z

Abstract. **Background**. Transposable elements (TEs) can represent one of the major sources of genomic variation across eukaryotes, providing novel raw materials for species diversification and innovation. While considerable effort has been made to study their evolutionary dynamics across multiple animal clades, molluscs represent a substantially understudied phylum. Here, we take advantage of the recent increase in mollusc genomic resources and adopt an automated TE annotation pipeline combined with a phylogenetic tree-based classification, as well as extensive manual curation efforts, to characterize TE repertories across 27 bivalve genomes with a particular emphasis on DDE/D class II elements, long interspersed nuclear elements (LINEs), and their evolutionary dynamics. **Results**. We found class I elements as highly dominant in bivalve genomes, with LINE elements, despite less represented in terms of copy number per genome, being the most common retroposon group covering up to 10% of their genome. We mined 86,488 reverse transcriptases (RVT) containing LINE coming from 12 clades distributed across all known superfamilies and 14,275 class II DDE/D-containing transposons coming from 16 distinct superfamilies. We uncovered a previously underestimated rich and diverse bivalve ancestral transposon complement that could be traced back to their most recent common ancestor that lived about 500 Mya. Moreover, we identified multiple instances of lineage-specific emergence and loss of different LINEs and DDE/D lineages with the interesting cases of CR1-Zenon, Proto2, RTE-X, and Academ elements that underwent a bivalve-specific amplification likely associated with their diversification. Finally, we found that

this LINE diversity is maintained in extant species by an equally diverse set of long-living and potentially active elements, as suggested by their evolutionary history and transcription profiles in both male and female gonads. **Conclusions**. We found that bivalves host an exceptional diversity of transposons compared to other molluscs. Their LINE complement could mainly follow a "stealth drivers" model of evolution where multiple and diversified families are able to survive and co-exist for a long period of time in the host genome, potentially shaping both recent and early phases of bivalve genome evolution and diversification. Overall, we provide not only the first comparative study of TE evolutionary dynamics in a large but understudied phylum such as Mollusca, but also a reference library for ORF-containing class II DDE/D and LINE elements, which represents an important genomic resource for their identification and characterization in novel genomes.

Towards a time-tree solution for Branchiopoda diversification: a jackknife assessment of fossil age priors.

Niccolò Righetti[1*], Filippo Nicolini[2*], Giobbe Forni[2], Andrea Luchetti[2]

[1] *Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Sorbonne Université, CNRS, IBPS, UMR7238, Paris, France.*

[2] *Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy.*

* the authors equally contributed to this work.

Submitted for peer-review.

Abstract. An understanding of Branchiopoda's evolutionary history is crucial for a comprehensive knowledge of the Pancrustacea tree of life, given their close evolutionary relationship with Hexapoda. Despite significant advances in molecular and morphological phylogenetics that have resolved much of the branchiopod backbone topology, a reliable temporal framework remains elusive. Key challenges include a sparse fossil record, long-term morphological stasis, and past topological inconsistencies. Leveraging a Bayesian Inference approach and the most extensive phylogenomic dataset for branchiopod to date, encompassing 46 species and over 130 genes, we inferred a time-calibrated phylogenetic tree. Furthermore, to strengthen the confidence in our divergence times estimation, we assessed the impact of age priors, topological uncertainties, and gene trees which are discordant from the species trees. Our results are largely consistent with the fossil record and with previous studies, indicating that Branchiopoda originated between 400 and 500 million years ago, and the orders of large branchiopods diversified during the Mesozoic. Concerning Cladocera, results remain problematic, with a sharper uncertainty in the diversification time with respect to the fossil record. Though, the jackknife resampling of fossils and the other sensitivity analyses proved our calibration method to be robust, suggesting that the difficulties in obtaining a paleontological-consistent time tree may be hindered by the variability in branchiopod substitution rates and topological instability within certain clades.