Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE DELLA TERRA, DELLA VITA E DELL'AMBIENTE

Ciclo 34

Settore Concorsuale: 05/B2 - ANATOMIA COMPARATA E CITOLOGIA

Settore Scientifico Disciplinare: BIO/06 - ANATOMIA COMPARATA E CITOLOGIA

GENES INVOLVED IN GERMLINE REGULATIVE PATHWAYS IN METAZOA: AN INTEGRATIVE APPROACH TO INVESTIGATE A KEY FEATURE OF ANIMAL BIOLOGY

Presentata da: Giovanni Piccinini

Coordinatore Dottorato	**Supervisore**
Maria Giovanna Belcastro	Liliana Milani

Esame finale anno 2022

Abstract

In Metazoa, the germline represents the cell lineage devoted to the transmission of genetic heredity across generations. Its functions intuitively evoke the crucial roles that it plays in the development of a new organism and in the evolution of the species. Germline establishment is tightly tied to animal multicellularity itself, in which the complex differentiation of cell lineages is favoured by the confinement of totipotency in specific cell populations. In the present thesis I addressed the subject of germline characterization in animals through different approaches, in an attempt to cover different sides and scales.

The first chapter of the thesis starts from the observation that the expression of many genetic elements is shared in germline-related lineages of all animals. Through transcriptomic analyses of online-available data, the extent and the nature of molecular signatures consistently upregulated in different species was investigated. In our data set, we could observe more frequently that the proportion of genes shared with other phyla among germline-related upregulated transcripts of a species was higher than expected by chance, confirming the less probable involvement of novel molecular factors in such mechanisms. When looking at the specific nature of elements involved, signals related to proper DNA replication resulted the most common across the considered species, while the regulation of transcription and post-transcriptional mechanisms appeared more variable, supporting for them a higher level of lineage-specificity.

On the other hand, the second chapter focuses on the molecular evolution and on the patterns of loss/expansion of a specific gene family, encoding for Tudor domain containing proteins. In animals, such family underwent novel evolution of many components that are tightly associated to germline-related pathways. Here, its evolutionary trajectories were investigated in a data set comprising 17 animal phyla. While the evolution of the Tudor domain could not be entirely outlined due to resolution issues, lineage-specific losses and expansions of the family were related to peculiar genomic dynamics and to the deep level of involvement of such proteins in the germline-associated piRNA pathway of retrotransposon silencing.

Lastly, in the third chapter, the characterization of the Tudor protein TDRD7, a Lotus and Tudor domain containing protein, was performed in the clam *Ruditapes philippinarum*, a species with annual gonad renewal. The protein expression in gametogenic animals retrieved similar results to the localization patterns of the germline marker *vasa*, previously characterized in the same species. This, together with functional data in other animals that support a Lotus-Vasa interaction, suggested that, in *R. philippinarum*, TDRD7 could be involved in the assembly of germ granules, i.e. cytoplasmic structures of variable complexity and appearance times that are related to germline specification and differentiation in virtually all animal germ cells, but whose assemblers can be taxon specific

Table of Contents

Preface 1

Chapter A – Looking for Germline-Specific Functional Signatures in Metazoa: an RNA-Seq Approach 5

Introduction 5

Materials and Methods 9

- Data set 9
- Transcriptome assembly and Differential Expression 11
- Comparative analyses 12

Results 13

- Species-specific Differential Expression analysis 13
- Shared germline-related homologous sequences: high representativeness of DNA replication-related genes 16
- DNA-related functions and domain-specific biases in germline-related samples 20

Discussion 23

- Considerations about the reliability of the samples and the analytical approach 23
- Germline-related genes are more frequently conserved across Metazoa 24
- The opposite trend might suggest lineage-specific adaptations 26
- Consistently upregulated OGs are coherent with germline-related totipotent processes and biased toward proper DNA replication 29
- Conclusions 32

Chapter B – The Tudor Domain Family: Shaping Factors Involved in the Evolution of Tudor Protein in Animals 33

Introduction 33

Materials and Methods 38

- Proteomes 38
- Identification of Tudor proteins and Tudor domain extraction 39
- Phylogenetic tree inference 40
- Prediction and evolution of the Tudor domain secondary structures 41
- Statistical analyses 42

Results 43

- Distribution of Tudor proteins in homology clusters 43
- Tudor protein distributions across species and statistical correlations with genomic features 44
- Phylogenetic inference of the Tudor domain tree 49
- Evolution of the Tudor domain secondary structures 52

Discussion 55

- Step-wise accumulation of N-t secondary structure could not be confidently confirmed, but the model remains a strong hypothesis 55
- G0 and G1 group distributions are partially explained by differences in genomic architecture 57
- The evolutionary patterns of G2 Tudor genes are consistent with multicellularity-related molecular innovations in animals 58
- *Piwi* loss and genomic dynamics as two driving forces of Tudor gene evolution: the cases of Tudor gene set reductions 60
- *Piwi* loss and genomic dynamics as two driving forces of Tudor gene evolution: the cases of Tudor gene set expansions 64
- Conclusions 67

Chapter C – Germline Differentiation in Bivalves: TDRD7 as a Candidate Factor Involved in *Ruditapes philppinarum* Germ Granule Assembly **68**

Introduction **68**

Materials and Methods **71**

- Sequence identification and analyses 71
- Sampling 73
- Primary antibodies 73
- Immunolocalisation 74
- Western blot 76

Results **77**

- TDRD7 in *Ruditapes philippinarum* 77
- *Tdrd7* upregulation in the gonads 77
- Western blot results for the different antibodies are not consistent 78
- Histological organization of the *Ruditapes philippinarum* gonad during the two stages of the reproductive cycle 80
- TDRD7 detection by immunofluorescent (IF) assay 81
- TDRD7 detection by immunohistochemistry (IHC) assay 84

Discussion **89**

- A candidate protein involved in germline differentiation 89
- Four antibodies, some different results: which one to trust? 90
- TDRD7 is localised in putative undifferentiated germ cells in the intestinal epithelium 93
- TDRD7 is expressed at different stages of gamete differentiation 94
- TDRD7 might be involved in the proper assembly of *Ruditapes philippinarum* germ granules 95
- Conclusions 96

Final Considerations 98

Bibliography 99

Supplementary Material 120

- Chapter A 120
- Chapter B 138
- Chapter C 144

Appendix – Mitonuclear Coevolution, but not Nuclear Compensation, Drives the Evolution of OXPHOS Complexes in Bivalves 148

Preface

Complex multicellularity, defined as obligate multicellularity characterized by different cell types and by a definite and regulated morphology, evolved independently at least five times in Eukaryota (Knoll 2011; Nagy 2017). One of the key features of such multicellular organisms is the subdivision of cellular mechanisms in different cell lineages, that morphologically and molecularly diversify. In the case of Metazoa, the scientific debate around the origin of multicellularity flowered almost two centuries ago and is still open nowadays (Brunet and King 2020). Despite metazoan phylogenetic relationships to other closely related unicellular eukaryotes are so far well characterized and solid (Ruiz-Trillo et al. 2008; Torruella et al. 2015), which were the first steps of the evolution of metazoan multicellularity is still matter of vivid discussion. Different models are ascribable in two main categories (Brunet and King 2017): either multicellularity preceded cell differentiation, that evolved later by division of labour through functional segregation (Arendt 2008); or cell differentiation was already present before the origin of multicellularity, and what evolved was a transition from a temporal succession of the various cell states to a spatial distribution of them within a colonial organism (Mikhailov et al. 2009; Sebé-Pedrò et al. 2017; Sogabe et al. 2019).

Despite the evolutionary steps that led to the establishment of animal multicellularity, it is evident that the last common ancestor of Metazoa already evolved many of the features related to it, since they are present in all extant species: the so-called Urmetazoa were most likely bacterivorous multicellular organisms with a proto-epithelium including collar cells, able to differentiate cells in various somatic states and in anisogamic germ cells (Richter and King 2013; Brunet and King 2017). Indeed, inseparably tied to the diversification of cell lineages is the existence of some cells that retain the whole potential of the organism cell states and that are devoted to the transmission of the genetic heredity across generations. These cells are the germ cells, that eventually produce gametes in sexual animals, and their lineage is called the germline.

This cell lineage was a key feature for the evolution of multicellularity in Metazoa because it allowed cells within the same organism to cover diversified roles without the burden of the transmission of the genome to the progeny. Some argued that this separation of roles, or rather the loss of totipotency in most differentiated cells, was itself the first and necessary step which allowed for the wide adaptive diversification of the somatic lineages observed in animals (Woodland 2016). Indeed, once the germline is established in an organism, all somatic cells become evolutionary dead-end, and any newly arisen mutation is doomed to be extinguished with the death of the individual. Germ cells, on the other hand, are kept in a totipotent state, carrying the Load and the Gold of genetic inheritance.

When and how this separation occurs within an organism are crucial features of the development of the organisms, and the mechanisms that fulfil this specification have been of interest since the XIX century. The initial definition of the germline by Weissman (1892) also included the concept of continuity for such cell lineage: indeed, the author defined a specialised physical component, called germ plasm, that was selectively inherited from the egg/zygote to the precursors of germ cells and was able to determine their fate. However, it was quickly assessed that this is just one of the two main modes by which germline specification can occur in animals. Indeed, species that display such mechanism of maternally deposited material are defined as having a germline specification through "preformation", while when the germ cell fate is determined later in embryogenesis by means of inductive signals from surrounding cells is defined as "epigenesis". Despite most of the classic model animals display a preformation mode of germline specification, it is now believed that epigenesis represents the ancestral phenotype, and that evolution of the former happened independently in many occurrences (Extavour and Akam 2003). Also, the convergence of preformation has been associated to higher evolvability and higher speciation rates, at least in vertebrates, due to the promptly "release" from totipotency that allows somatic lineages to diversify earlier during development (Evans et al. 2014; Johnson et al. 2015), somehow in resonance with the Woodland (2016) model of early germline segregation as the adaptive driver of complex multicellularity.

The molecular revolution of the recent decades allowed to emancipate the germ cell investigation from exclusively morphological description, and it was possible to determine the genes involved in the determination and differentiation of various animal germlines. One of the most interesting results related to such molecular characterization was the observation that some of the dedicated genes were the same in virtually all studied species. Moreover, it was also possible to assess that these molecular signatures were not exclusively associated to the germline, but they were present, in some animals, also in different stem cell lineages that comprehended also somatic potential. This led to associate the programming of germ and multipotent cells thanks to the shared expression of a highly conserved genetic toolkit (Juliano et al. 2010).

The subsequent theoretical step was the expansion of the definition of germline so that it would also comprehend these bivalent germ/soma lineages, therefore including all cells capable to give rise to a germ cell, at least potentially (Solana 2013). This new definition of germline rejuvenated the Weissman theory of germline continuity without the necessity to invoke an uninterrupted inheritance of cytoplasmic structures, but considering continuity as the steady expression of a homologous genetic toolkit. What is now clear is therefore the tight association of many totipotent lineages in animals, delineating a thread for the evolution of germline itself. Indeed, the fact that germline genes were observed as expressed in totipotent somatic cell lineages, like sponge archaeocytes, cnidarian interstitial cells, and planarian neoblasts, has provided fertile suggestions both for the evolution of the genetic elements involved, and for considerations about the separation of cell lineages at the origin and during the early developing of multicellularity.

The first chapter of the present PhD thesis (Chapter A) is pertinent to these topics and represents an attempt to outline transcriptional similarities across the germlines of different species. By analysing online available data, a selection of RNA-Seq experiments including germline-related and somatic samples was performed in order to have a widespread phylogenetic species coverage. Assessing the species-specific transcriptional profiles and comparing the obtained signals across the different species, allowed to delineate whether shared mechanisms were present without the investigation on an *a priori* set of determined genes. This analysis covered a data set sampling of wide phylogenetic scale, comprising 10 species belonging to 8 phyla, spanning from classic model organisms to early branching clades, and investigated around the whole set of transcription profiles, in the attempt to uncover shared aspects.

In the second chapter (Chapter B), while keeping a metazoan-level scale, the evolution of a specific protein family is investigated. Indeed, the establishment of multicellularity in animals was coupled to both the expansion and co-option of previously evolved genes, and the evolution of completely novel ones, whose homologues so far have not been found in other eukaryotes (Richter and King 2013; Brunet and King 2017). For instance, the protein families of Wnt and TGF-β, with key roles in morphogenesis and developmental cell fate, are metazoan novelties, while the numerous Tyrosine Kinases involved in cellular signalling are the product of a metazoan-specific family expansion (Richter and King 2013; Brunet and King 2017). Also germline differentiation, as said before, was a crucial step during the establishment of multicellularity, and genes related to it followed similar patterns of evolution, with cases of co-option, expansion, and novel evolution. For instance, key germline genes like *vasa* and *piwi* represent metazoan novelties, while others were already present, at least in Holozoa (Fierro-Constain et al. 2017). Proteins harbouring the Tudor domain are an example of animal germline-related family expansion: indeed, many Tudor proteins are present in most Eukaryota, but a numerous set of them is the product of a recent metazoan gene radiation, whose products are usually associated to the germline-specific piRNA pathway of retrotransposon silencing, representing a highly interesting investigational unit. An analysis regarding the evolutionary pathways that the Tudor domain protein family underwent in Metazoa through reassemblies, losses and expansions, is examined in the second chapter of the present thesis, that covers the domain molecular evolution in 93 metazoan species comprising 17 phyla. The choice of this specific protein family derived from the interrelated study described in Chapter C.

Indeed, the third and last chapter (Chapter C) focuses on a narrower scale, both at the target and at the taxonomic level. In particular, the expression profile of the Tudor-Lotus domain-containing protein TDRD7 is analysed in a target species, proposing it as a putative element involved in the assembly of germline-related cytoplasmic structures, i.e. germ granules. Starting from literature data of known germ granules assemblers, we identified by bioinformatic investigation such protein as a candidate factor covering similar functions during the germline differentiation in the clam *Ruditapes philippinarum* (Mollusca Bivalvia). The characterization of TDRD7 in such species is performed both at the transcription level *in silico* and at the expression level by immunolocalization in tissue. This species model is interesting for many aspects and has been widely studied due to its peculiar mitochondrial inheritance mechanism. However, the feature that is taken advantage of in the present thesis is the annual renewal of the gonads, that sees the reprise of germline differentiation pathways at each reproductive season. Previous works assessed the involvement of the classic germline marker gene *vasa* in the specialisation of germ cells of this species, identifying clusters of cells within the intestinal epithelium as putative repositories of undifferentiated germ cells (Milani et al. 2015, 2018). Moreover, germ granules associated to meiosis onset, and possibly to preformation mechanisms, have been observed (Reunov et al. 2019), invoking the need for a better understanding of the molecular features involved. Given the key roles covered by germ granules in virtually all animal germ cells, a deep understanding of their processing is helpful for grasping their nature and the different declensions in which they can be observed.

Chapter A

Looking for Germline-Specific Functional Signatures in Metazoa: an RNA-Seq Approach

Introduction

Germ cells have been the subject of scientific interest for centuries, but until few decades ago, their identification and study were guided only by cytological observations of typical morphological features, such as their high nucleo-cytoplasmic ratio and the presence of cytoplasmic granular bodies usually in a perinuclear position (see Chapter C). With the advent of molecular technologies, the molecular factors involved in differentiation and specification of the germline started to be identified. However only with the advancement of such technologies in recent times it became possible to characterize the genetic networks and the protein profiles of germ cells in a still expanding number of animal species, allowing to better delineate and define the characteristic of a cellular lineage that represents one of the most important and ancestral features of animal multicellularity.

One of the most interesting observations that could be made about the molecular profiles of germ cells was that they are characterized by the shared expression of a gene set that is highly conserved among animals (Ewen-Campen et al. 2010; Juliano et al. 2010; Fierro-Constaìn et al. 2017). Transcription and expression of some of these gene have been observed in virtually all animals in which molecular germline characterization has been performed (see references within Extavour and Akam 2003 and Fierro-Constaìn et al. 2017) and are usually associated with post-transcriptional regulatory activities.

For instance, one of the most typical molecular signatures of animal germ cells is the transcription and expression of homologues of the *vasa* gene (Juliano et al. 2010; Lasko 2013). This gene encodes for a DEAD-box RNA helicase that acts as a translational activator with sequence-specific activity (Liu et al. 2009; Gustafson and Wessel 2010; Lasko 2013). It was first associated with the specification of Primordial Germ Cells (PGCs, i.e. the first cells with exclusively germ fate) in *Drosophila melanogaster* embryos, them being absent in mutants of the gene (Shupbac and Wieschaus 1986), but later was associated to germline formation in animals in general, with functions that span from RNA regulation, including the aforementioned selective mRNA translation promotion (but also with roles in the piRNA pathway of Transposable Elements – TEs – silencing), to chromatin condensation during female germline mitosis (functions reviewed in: Lasko 2013).

Another key molecular factor associated with germline is the *piwi* gene. Such gene belongs to the Argonaute protein family and, together with its close homologues *aub* and *ago3* (according to *D. melanogaster* nomenclature), it is involved in the germline-related piRNA pathway. Also for this gene, the first associations with proper germ cell formation were obtained in model organisms, namely *D. melanogaster* and *Caenorhabditis elegans* (Cox et al. 1998), but later its expression was observed in the germline of animals spanning the whole metazoan phylogenetic tree, from ctenophores to annelids (Lim and Kai 2015). The functions of Piwi are strictly related to piRNA-mediated RNA silencing, mostly focused on retrotransposon silencing (Juliano et al. 2011; Ku and Lin 2014; Czech et al. 2018). Mature piRNAs bind to Piwi, and the resulting complexes are able to bind cytoplasmic mRNAs belonging to retrotransposons, avoiding their translation and leading them to degradation, and are also able to enter the nucleus and act as transcriptional suppressors through epigenetic mechanisms (Lim and Kai 2015; Czech et al. 2018). The germline-specific expression of *piwi* is therefore necessary for genome homeostasis in this totipotent cellular lineage devoted to genetic inheritance, avoiding massive replication and mobilization of retrotransposons.

A third genetic element that has been widely associated with the germline is *nanos*. First described in *D. melanogaster* as a molecular determinant for the formation of embryonic posterior region, it was then directly associated to the differentiation of functional germ cells (Kobayashi et al. 1996). Also in this case, it was later related to the same functions in most other animals (Fierro-Constaìn et al. 2017). Homologues of *nanos* in Metazoa encode for proteins with different sequence composition, but they all share a typical, widely conserved, C-terminal Zinc-finger domain of the CCHC type. This domain has RNA-binding activities (Hashimoto et al. 2010), and thanks to the association with multiple other factors, Nanos proteins can bind a diverse set of mRNAs, controlling their translation fate usually in a repressive way, but with also some documented promoting activities (Keuckelaere et al. 2018).

Many other molecular factors have been associated to germ cell specification/differentiation in different animals through the years (for a review on the molecular machinery of germline specification see: Ewen-Campen et al. 2010), such as the RNA-binding translational activator Boule (Shah et al. 2010), the RNA-binding translational repressor Pumilio (e.g. Parisi and Lin 1999; Nakahata et al. 2003), the methyl-lysine/arginine readers belonging to the Tudor protein family (see Chapter B), the translational repressor germ-cell-less (Leatherman et al. 2002), or the post-transcriptional regulator and alternative mRNA splicing factor Bruno (Kim-Ha et al. 1995; Hashimoto et al. 2006). Among all these genes, however, *vasa*, *nanos*, and *piwi* are those that are mostly shared in the germline of different Metazoa, making them quasi-universal markers of germ cells, regardless of their differentiation stage: most other factors are indeed transcribed and expressed in specific germ cell stages, and/or they have not been associated to germline functions in all animals (see for instance the summary tables of germline determinants in: Extavour and Akam 2003; Ewen-Campen et al. 2010; Juliano et al. 2010; Fierro-Constaìn et al. 2017).

However, during the precise and wide characterization of germ cell molecular determinants in the last decades, it has been simultaneously assessed that such factors were not exclusively confined in the germline, but rather their expression was observed also in animal multipotent cell lineages that comprehended also somatic cells among their potential fates. For instance, in the sea urchin *Strongylocentrotus purpuratus*, the embryonic small micromere lineages accumulate *vasa*, *piwi*, and *nanos* mRNAs, but their differentiation potential encompass the totality of the adult tissues, and not strictly germ cells (Juliano et al. 2006). Also, in mollusc embryos, the 4d blastomere expresses germline determinants, despite its progeny not being limited to the germline (Kranz et al. 2010), exactly like in the cells of the mesodermal posterior growth zone of annelid embryos, that gives rise to both somatic and germ stem cells (Rebscher et al. 2007), showing that the presence of such genetic factors precedes the actual determination of strict germ cell fate.

Moreover, the expression of germline determinants is not limited to embryonic stem cells, but is present also in adult stem cells of some animals. In the aforementioned Annelida, also stem cells involved in the posterior elongation during post-caudal regeneration of adults share some germline-associated factors (Gazave et al. 2013), together with cells involved in body regeneration and asexual reproduction (Tadokoro et al. 2006; Ozpolat and Bely 2016). In *Hydra* adult specimens (Cnidaria, Hydrozoa), *vasa* and *nanos* transcripts are present both in germ cells and in interstitial cells, that is a multipotent cell population that can give rise to both germ cells and different kinds of somatic cells, namely nematocystes, neurons, and glands (Mochizuki et al. 2001). In *Ephydatia fluviatilis* and *Amphimedon queenslandica* (two Porifera of the class Demospongiae), homologues of *piwi*, *vasa*, and *nanos*, are expressed in the adult stem cell system, that is composed of the two cellular lineages of choanocytes and the totipotent archaeocytes, both capable of self-renewal and gamete production (Funayama et al. 2010; Funayama et al. 2013; recent findings also expand such observations to Homoscleromorpha sponges, and precisely to their archaeocyte-like type 2 vacuolar cells: Fierro-Constaìn et al. 2017). Again, in adult flatworms (Platyhelminthes), the tremendous whole-body regenerative capabilities are due to the extensive and diffused presence of totipotent stem cells called neoblasts, that express many of the germline-related genetic toolkit (reviewed in: Krishna et al. 2019). Neoblasts are totipotent cells that were observed also in the early-branching bilaterian Acoela, and also in these cells transcription of *piwi* and *vasa* is present (Gehrke and Srivastava 2016). Lastly, also in Tunicata (Chordata), whole-body regeneration capabilities have been associated to *piwi* expression in cells of the internal epithelium of blood vessels (Rinkevich et al. 2010).

Altogether, these observations suggest a broad molecular similarity between germ cells and stem cells, leading to the theorizing of a Germline Multipotency Program (GMP), i.e. of a genetic toolkit that operates both in the germline and somatic multipotent stem cell lineages and that is fundamental for establishing and maintaining multipotency (Juliano et al. 2010). Later, Solana (2013) synthesized two centuries of germline-associated morphological and molecular studies by proposing the definition of Primordial Stem Cells (PriSCs). According to Solana, these cells are highly conserved stem cells that basically include all stages that exist between the zygote and the first specified cells with exclusive germ cell fate (i.e. PGCs). The author proposed these PriSCs, despite their mixed germline-somatic potential, to be included into the germline, that would then comprise all cells potentially capable of producing a germ cell. By defining such kind of cells, all controversies regarding the continuity of the germline throughout generations were solved (re-establishing under a new light the famous Weissman's barrier), since, in previous years, classical definitions of germline were biased toward sexually reproducing animals. For instance, all aforementioned examples of stem cells that had both somatic and germline potential (interstitial cells, archaeocytes, neoblasts, mesodermal posterior growth zone cells, etc.), that previously would not have fall into the classical definition of germline, can now be considered PriSCs, establishing a continuity from zygote to germ cells. Therefore, the definition of totipotent PriSCs both solve germline continuity controversies, and collect within the same definition, and perhaps within the same homologous lineage, totipotent cell lineages.

All these findings highlight the crucial role that the GMP genes had in animal evolution and suggest that the most recent common ancestor of all animals already had most of them. Interestingly, while the evolution of most of these genes predated the separation of the animal lineage from other eukaryotes, *vasa*, *nanos*, and *piwi* (together with some strictly germline-related Tudor proteins; see Chapter B) are thought to be specific metazoan innovations. Indeed, so far homologues have not been found in any other eukaryotic lineage, differently from other GMP determinants that have been annotated at least in other holozoan (e.g. *bruno*, *pumilio*, and *boule*; Alié et al. 2015; Fierro-Constaìn et al. 2017).

Therefore, both the evolution of novel molecular factors and the co-option of other ancestral genetic elements led to the evolution of the GMP, that was a crucial step in the evolution of animal multicellularity itself. Indeed, the origin of multicellularity has been interpreted as guided by the selective advantage of differentiation of various cell lineages and the distribution of biological functions. Woodland (2016) proposed that the very first of these differentiations was indeed the separation between the germline (comprising PriSCs) devoted to reproductive functions and cell renewal, and the somatic lineages, that in this way were free to diversify.

In the present analysis we aimed to explore the transcriptional signatures of germline-related tissues and cell lineages in different animals. We took advantage of online available experimental data to retrieve as much RNA-Seq experiment as we could that fit the established features of having enough samples size to assess transcript abundance, and of having control somatic samples produced within the same experiment. We performed species-specific differential expression analyses, and we then checked whether there were some homologous genes consistently upregulated in the germline-related samples for most of the species, in order to retrieve a common transcriptional signal that could have emerged despite the data set heterogeneity. Moreover, using reference proteomic data from other animals, we looked into the upregulated species-specific germline-related transcripts to get hints on how many of them were represented by lineage-specific innovations.

Materials and Methods

Data set

All RNA-Seq reads used in the present study were downloaded from the Short Reads Archive of NCBI (https://www.ncbi.nlm.nih.gov/sra). We searched for female germline-related samples (i.e. the lineage that maintains totipotency throughout development: Seydoux and Braun 2006) in metazoan RNA-Seq experiments generated through Illumina platforms with the following key-words: oocyte(s), gonad(s), egg(s), germline, germ line, germ cell(s). The search results were then filtered for experiments that included both samples belonging to exclusively germline-related tissues or cells and also any kind of somatic tissue within the same project, and contemporarily for experiments that included at least 2 biological replicates for condition. We then chose the final data set keeping an even representativeness among taxa.

The candidates belonged to 11 species: *E. fluviatilis* (Porifera), *Clytia hemisphaerica* (Cnidaria), *Brachionus manjavacas* (Rotifera), *C. elegans* (Nematoda), *Danio rerio* (Chordata), *Xenopus tropicalis* (Chordata), *D. melanogaster* (Arthropoda), *Penaeus chinensis* (Arthropoda), *Ruditapes philippinarum* (Mollusca), *Haliotis rufescens* (Mollusca), and *Eisenia fetida* (Annelida). From these, *E. fetida* was excluded because the germline-related samples were represented by whole bodies enriched for gonads, and not only the specific tissue of interest. Also *P. chinensis* was excluded during the analyses because an over-representation of stress-related signals emerged during the Differential Expression analysis, invalidating the confidence of the samples. We also decided to include among our samples RNA-Seq reads of *Schmidtea mediterranea* neoblasts (and differentiated progeny as somatic control). These cells, together with multipotent cells of other Metazoa, have been associated to the germline since neoblasts express germline-associated signature genes, leading to theorise the existence of the GMP shared by totipotent germ cells (see Introduction; Juliano et al. 2010; Solana 2013). Therefore, the final data set comprehended 10 species covering 8 phyla (Table 1).

This data set was extremely heterogeneous in sample composition (see Table 1), an unavoidable flaw of using online available data that were not originally intended for such comparative analyses. Indeed, this study was intended as a pilot study before planning specific sequencings from different phyla in a more standardized way: analyses that are surely of our interest, but that will require much more time, funding and collaboration efforts from different research groups. However, we are convinced of the reliability of the present analysis, despite its inherent limits. Both the fact that the somatic controls belonged to different non-homologous tissues, and the fact that what we call germline-related samples were whole gonads for some species and cell populations for others, did not compromise the principles of the analysis, but rather its power. The heterogeneous nature of our data set might have prevented a strong signal to emerge, but if something could be observed, it meant that a shared signal was indeed present in the least common multiple of all samples, that is the germline. In other words, we are convinced that our study was not subjected to the risk of observing false positives, but, rather, to the risk of having a great amount of false negatives.

Table 1. Sample composition of the 10 species included in the data set. The number of replicates for each sample represents biological replicates.

Species	Phylum	BioProject (NCBI database)	Germline-related samples (n° replicates)	Control somatic samples (n° replicates)
Brachionus manjavacas	Rotifera	PRJNA345262	Eggs (2)	Whole bodies with eggs removed (2)
Caenorhabditis elegans	Nematoda	PRJNA392422	Embryonic Primordial Germ Cells (3)	Embryonic somatic cells (3)
Clytia hemisphaerica	Cnidaria	PRJNA393679	Growing oocytes (2)	Somatic gonadic endoderm/ectoderm (4)
Danio rerio	Chordata	PRJEB30097	Gonads (2)	Livers (2)
Drosophila melanogaster	Arthropoda	PRJNA388952	Gonads (4)	Genitalia (4)
Ephydatia fluviatilis	Porifera	PRJNA244851	Archeocytes (2)	Mixed differentiated cells (2)
Haliotis rufescens	Mollusca	PRJNA488641	Gonads (2)	Mantles (2)
Ruditapes philippinarum	Mollusca	PRJNA672267	Gonads (8)	Mantles (8)
Schmidtea mediterranea	Platyhelminthes	PRJNA503908	Neoblasts (3)	Mixed differentiated cells (3)
Xenopus tropicalis	Chordata	PRJNA381064	Gonads (2)	Hearts and livers (4)

Transcriptome assembly and Differential Expression

Given that RefSeq genomes were not available for all the species of our data set, we decided to uniform any kind of computational bias among our samples and we performed a *de novo* transcriptome assembly for all. Assemblies were performed for each species with Trinity v2.9.0 (Grabherr et al. 2011) by pooling all samples together, with default parameters for read normalization. Read quality filter was performed with Trimmomatic v0.39 (Bolger et al. 2014) using a sliding window size of 1/5 of the read length with a cut-off phred score of 28, and excluding all reads shorter than 2/3 of read length.

To reduce complexity, we collapsed transcripts through CD-HIT v4.8.1 (Li and Godzik 2006) at 99% of identity. We then filtered the transcriptomes by keeping exclusively transcripts that had a metazoan best hit as result of a DIAMOND v2.0.6.144 search (Buchfink et al. 2021) against the non-redundant protein database of NCBI (10^{-5} e-value cut-off). The completeness of the filtered transcriptomes was evaluated through the BUSCO v5 set of core metazoan orthologues as implemented in the gVolante website (https://gvolante.riken.jp/index.html).

Since we were interested in Coding Sequences (CDSs) only, we also performed an Open Reading Frame (ORF) prediction through TransDecoder v5.5.0 (https://github.com/TransDecoder), keeping the single best ORF for each transcript. To help inferring the most likely ORF position within the transcript, the software was also fed with a DIAMOND search against Swiss-Prot (10^{-5} e-value cutoff; The UniProt Consortium 2021) and an HMMscan (HMMER v3.2.1; Eddy 2011) against Pfam-A (Mistry et al. 2021). Only transcripts with a predicted ORF were considered for the subsequent analyses. From now on, we will refer to the "translated ORFs" as "translated transcriptomes", and when we refer to "transcripts" we mean "ORF-including transcripts", i.e. those supposedly belonging to protein-coding genes.

Transcript quantification was performed for each species with perl scripts included in the Trinity utilities package, with Salmon v1.3.0 (Patro et al. 2017) as the tool for expression estimates. Differential Expression (DE) analyses were then performed with both DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010), as implemented in the Trinity utilities package. We decided to moderately raise the strictness for the DE analysis: only transcripts with a $\log_2$ fold change higher than 1 in the germline-related samples (i.e. twice as abundant in respect to the control somatic samples), with a corrected p-value lower than 10^{-3} and significant for at least one analytic tool, were considered as differentially upregulated.

For each species set of upregulated transcripts, we wanted to calculate the proportion of sequences that shared homology across Metazoa and the proportion of lineage-specific ones, i.e. a phylostratigraphic analysis of the germline-related upregulated transcriptomes. To do that, we downloaded 111 proteomes from online databases (covering 21 animal phyla, comprehending all those belonging to our data set species, and 4 unicellular holozoan phyla, i.e. the closest relatives to Metazoa; these proteomes were the same used in Chapter B for the Tudor domain analysis; see Supplementary Table B1) and ran a homology inference between them and our 10 species translated transcriptomes. The analysis was run with OrthoFinder v2.3.11 (Emms and Kelly 2019) with the --ultra-sensitive parameter (highest sensitivity) and all sequences that ended up within the same cluster (OrthoFinder's OrthoGroups, or OGs) were considered homologous. An upregulated germline-related transcript was considered inter-phyletic when it shared homology with at least another sequence outside the belonging phylum. If a CDS ended up within an OG composed exclusively of intra-phyletic sequences, we considered it phylum-specific. CDSs that shared at least one homologue outside the belonging phylum were considered metazoan-specific if no sequences of non-metazoa Holozoa were comprehended in their OG.

Comparative analyses

To observe whether there were any homologous CDSs consistently upregulated in different species of our data set, we first constructed clusters of homology for the whole translated transcriptomes of our 10 species. We used OrthoFinder with the same parameters for the phylostratigraphic analysis previously exposed. CDSs of different species were considered co-upregulated in germline-related samples between two species when they were significantly differentially transcribed (see previous sub-chapter) and belonging to the same OG. For OGs that comprehended sequences consistently upregulated in the majority of the species (8 or more out of 10), we specifically annotated the sequence content by BLAST searches based on the sequences of *C. elegans*, *D. rerio*, *D. melanogaster*, and *X. tropicalis*, since for these models the confidences of online annotations are high, and functional data are available. For other OGs, we counted the number of times that all possible combinations of species ended up within the same germline-related OGs. In this way we could count how many times each combination of species had a shared set of germline-related upregulated CDSs, and we calculated the deviation from expected random distributions with the UpSetR R package as implemented online (https://vcg.github.io/upset/; Conway et al. 2017).

We also ran InterProScan v5.45.80 (Jones et al. 2014) on the whole translated transcriptomes of all 10 species, annotating for each sequence the associated Gene Ontology (GO) terms and InterProScan (IPR) codes. We performed a GO term enrichment analysis (topGO package on R; Alexa and Rahnenfuhrer 2021) to observe which biological processes and molecular functions were significantly enriched in each species germline-related samples. Given the diversity of our data set and the cloudy nature of GO term enrichment analyses, we decided to look in a comparative manner only to the strongest signals emerged. We indeed considered within each species only those GO terms that were annotated in germline-related transcripts at least twice as much as randomly expected. We then looked at such germline-related enriched GO terms shared by at least 50% of our data set (i.e. at least in 5 species). Visualization of semantically similar GO terms was performed on the ReViGO server with a collapsing threshold SimRel value of 0.9 (http://revigo.irb.hr/; Supek et al. 2011).

A similar, but not overlapping, analysis was performed with IPR codes. For each IPR code of the InterProScan database (that are annotation codes corresponding to both domains, motifs, and protein families), we counted the species-specific number of CDSs respectively annotated in the germline-related subset and in the full translated transcriptome. When an IPR code was annotated exclusively among germline-related CDSs, we considered it as biased toward germline-related samples. For all other IPR codes, we tested whether they were significantly biased. We performed an odds ratio test (*odds.ratio* test in R, *questionr* package), that associate a p-value to the comparison of two ratios: the ratio of appearance of each IPR code in germline-related sequences was compared to the ratio in the whole translated transcriptome. A p-value lower than 0.05 for the test meant that the IPR code was biased toward germline-related samples. Comparative analysis were performed considering IPR codes that were biased in more than 50% of the species.

Results

Species-specific Differential Expression analysis

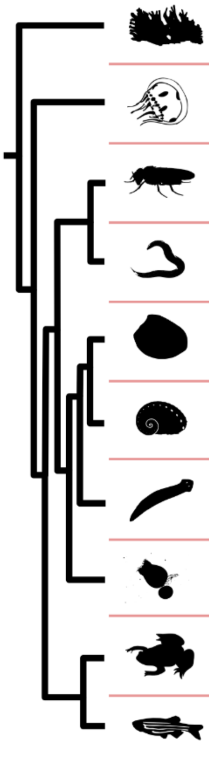
The total number of transcripts resulting from the *de novo* transcriptome assemblies and filtering are summarized in Figure 1, together with completeness statistics. The BUSCO quality check revealed high levels of completeness for most of the filtered transcriptomes, with a proportion of complete core genes never lower than 93%, and of complete+partial never lower than 95%. The exceptions were represented by *S. mediterranea*, *B. manjavacas*, *C. elegans*, and *E. fluviatilis*, that had lower completeness statistics but nevertheless not so low to invalidate subsequent analyses (complete+partial: 80.29%, 89.94%, 73.90%, and 90.15%, respectively). This could have been caused by lineage-specific diversification that prevented the detection of metazoan-wide core orthologues. To check for this possibility, we reran the completeness analysis for *C. elegans* with the Nematoda specific core gene set (the only one out of the four species for which a phylum-specific database was implemented by BUSCO), obtaining only slight improvements (80.10% complete+partial). We do not think that the lower completeness statistics were due to transcriptomes built *de novo* and not on reference genomes (see Materials and Methods), since for other species we obtained good results despite the same assembly method. Instead, lower values for these 4 species in respect to the other 6 transcriptomes could be due to the sample type, since in the former ones the samples were cell populations, while in the latter ones they were tissues (see Table 1). Indeed, it is more likely to miss transcription of some core gene in cell populations rather than in pools of different tissues that comprise diverse cell lineages. However, the levels of completeness were still relatively high, and the lower levels might have brought the subsequent analyses toward false negatives rather than false positives, therefore not invalidating the obtained result but at most limiting the detection power itself. On average, ~11% of each species translated transcriptome was differentially upregulated in germline-related samples in respect to somatic controls (twice as transcribed; p-value < 10^{-3}; Figure 1). The variability was consistent (50.6% coefficient of variability, calculated as the standard deviation over the mean) but it could be due to the heterogeneity in the sample tissue compositions between the different experiments. However, a much more interesting signal was represented by the percentages of phylum-specific germline-related CDSs (i.e. CDSs that did not share homology with any other sequence outside the belonging phylum; see Materials and Methods for details). Indeed, such statistics differed widely between species, passing from 4.6% for the cnidarian *C. hemisphaerica* to roughly 32% in the nematode *C. elegans* and the rotifer *B. manjavacas* (overall mean of 16.6%, with 64.6% of coefficient of variability; Figure 1).

However, by calculating the ratios between these percentages and the same kind of percentages calculated for the whole translated transcriptome but excluding those transcripts upregulated in germline-related samples, we could assess whether there was any bias toward intra-phyletic or inter-phyletic homology in the germline sequence subsets:

- If the ratio between the two percentages was lower than 1, then it would mean that it was more likely for a germline-related CDS to share homology with at least another sequence of another phylum.
- On the contrary, a ratio higher than 1 meant that for that species the germline-related subset had a higher proportion of phylum-specific CDSs in respect to the rest of the translated transcriptome (ratios are summarized in Figure 1).

Half of the data set had a ratio lower than 1, indicating a bias toward germline-related upregulation of shared inter-phyletic genes. However, 3 species, namely *S. mediterranea*, *B. manjavacas*, and *C. elegans*, displayed the opposite signal, with a higher percentage of phylum-specific germline-related transcripts in respect to the rest of the transcriptome (1.381, 1.344, and 1.170 phylum-specific ratios, respectively). These species were also 3 out of the 4 species that had low levels of transcriptome completeness as inferred by BUSCO. This however should not be an issue for this calculation, since the incompleteness of a transcriptome is usually due to experimental technical issue, that should not have any bias toward the selective maintenance of lineage-specific transcripts. Supporting this, *E. fluviatilis*, that had a BUSCO completeness percentage almost equal to that of *B. manjavacas*, displayed one of the lowest ratios between germline- and non-germline-related lineage-specific CDSs (0.456 phylum-specific ratio).

Moreover, a similar percentage ratio was calculated for all those sequences that were shared by at least two phyla. If these sequences did not have even one homologue outside Metazoa (8 Holozoa species covering 4 phyla were included in the analysis), they were considered as metazoan-specific. We could observe that the percentage of these metazoan-specific CDSs was lower in germline-specific samples with respect to the rest of the translated transcriptome for all species (excluding *B. manjavacas*; Figure 1), meaning that it was more likely for a holozoan-shared CDSs to be differentially expressed in germline-related samples.



Species	BUSCO completeness (complete + partial)	N° of CDSs	Germline related CDSs (% to whole)	Phylum specific germline related CDSs	Phylum specific ratio	Metazoa specific ratio
Ephydatia fluviatilis	87.84% (90.15%)	20267	2900 (14.309%)	7.345%	0.456	0.440
Clytia hemisphaerica	93.19% (95.18%)	17071	1941 (11.370%)	4.637%	0.422	0.678
Drosophila melanogaster	93.71% (96.44%)	14270	3325 (23.301%)	14.887%	1.081	0.811
Caenorhabditis elegans	69.81% (73.90%)	11659	1583 (13.577%)	32.244%	1.170	0.651
Ruditapes philippinarum	95.39% (98.22%)	54622	2992 (5.349%)	7.461%	0.394	0.539
Haliotis rufescens	96.75% (98.11%)	33202	1533 (4.617%)	10.633%	0.881	0.851
Schmidtea mediterranea	72.75% (80.29%)	79696	5123 (6.428%)	29.397%	1.381	0.878
Brachionus manjavacas	81.24% (89.94%)	28446	4017 (14.121%)	32.064%	1.344	1.348
Xenopus tropicalis	97.59% (98.85%)	36549	4348 (11.896%)	10.281%	0.448	0.681
Danio rerio	99.16% (99.27%)	44165	3081 (6.976%)	17.592%	0.969	0.832

Figure 1. Transcriptomic statistics. Phylogenetic relationships between the species are schematized on the left (referring to Laumer et al. 2019). **BUSCO completeness** is calculated on the whole transcriptome. **N° of CDSs** represents the number of transcripts for which an ORF could be extracted, i.e. Coding Sequences. **Germline-related CDSs** correspond to the number of ORF-containing transcripts that were upregulated in germline-related samples (the percentage is calculated on the whole set of ORF-containing transcripts). **Phylum-specific germline-related CDSs** corresponds to the percentage of upregulated germline-related CDSs for which not even one homologous sequence could be found outside the belonging phylum. **Phylum-specific ratio** is calculated as: the previous percentage over the phylum-specific percentage of all other CDSs of the transcriptome. **Metazoa-specific ratio** is calculated as the percentage of metazoan-specific CDSs in germline-related samples over the same percentage in all other transcripts (both percentages are calculated considering only those transcripts shared by at least two phyla, i.e. excluding phylum-specific CDSs). Phylum- and Metazoa-specific ratio lower than 0.9 are depicted in green; those higher than 1.1 are depicted in red.

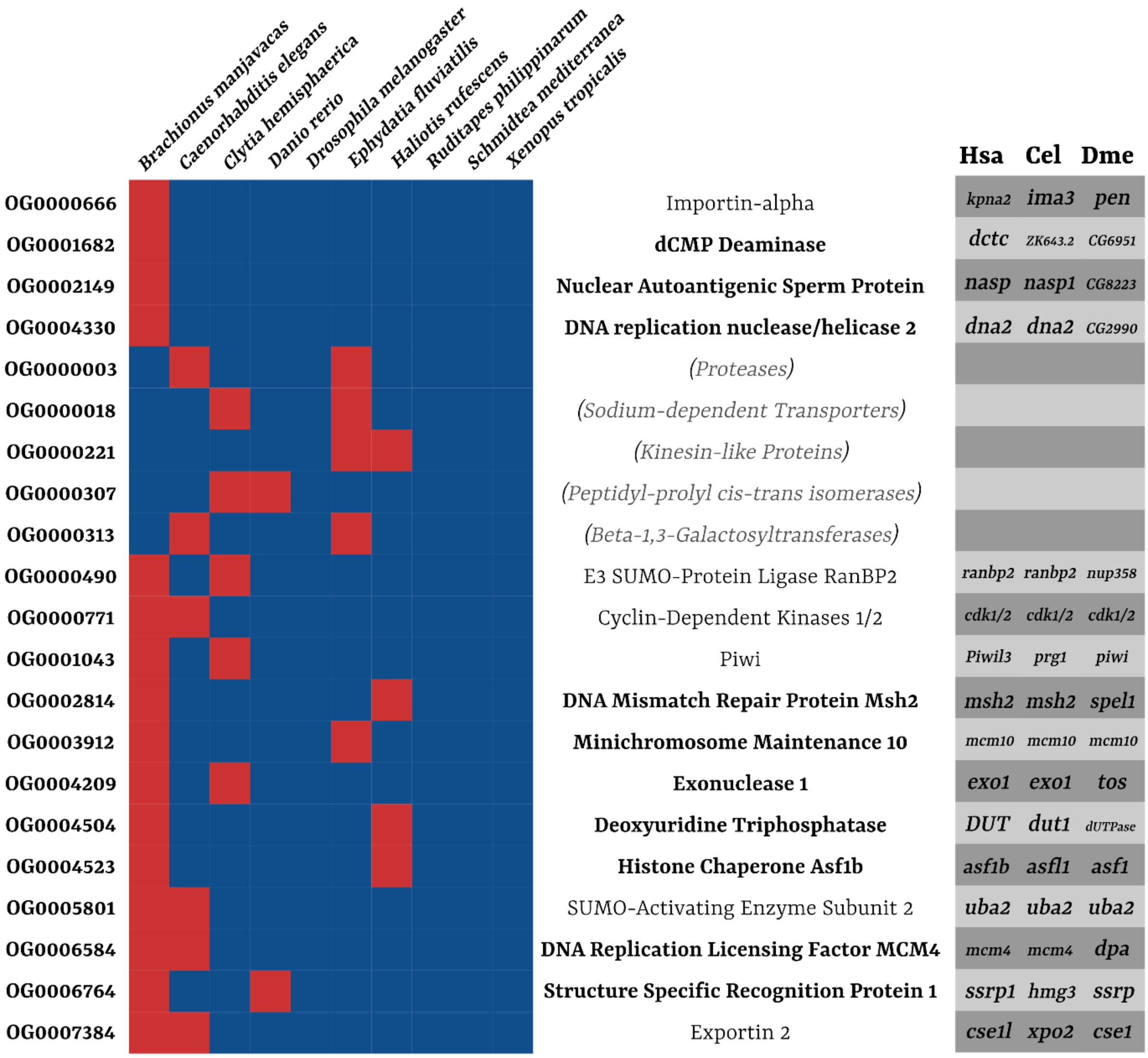
Shared germline-related homologous sequences: high representativeness of DNA replication-related genes

We identified 4365 OGs that included germline-related CDSs consistently upregulated in at least two species of our data set. Out of these, 4 OGs were consistently upregulated in 9 out of 10 species: in all cases it was *B. manjavacas* that lacked differential transcription of the homologous CDSs (Figure 2). These OGs included homologues encoding for Importin-alpha (one of the two subunits of Importin, involved in protein import inside the nucleus, but also in centrosome duplication and mitotic spindle dynamics), dCMP Deaminase (involved in nucleotide synthesis for DNA), the Nuclear Autoantigenic Sperm Protein (a histone-binding protein involved in DNA replication-dependent nucleosome assembly), and DNA replication nuclease/helicase 2 (predicted to be involved in proper DNA replication).

Other OGs for which we specifically annotated the content were the 17 ones with germline-related upregulated sequences shared by 8 species (for 12 out of these 17 OGs, *B. manjavacas* was missing, but the other missing species was variable; Figure 2). These OGs included 7 genes with activities directly related to DNA (especially DNA proper replication) that encoded for: DNA Mismatch Repair Protein Msh2, Minichromosome Maintenance 10, Exonuclease 1, Deoxyuridine Triphosphatase, Histone Chaperone Asf1b, DNA Replication Licensing Factor MCM4, and Structure Specific Recognition Protein 1. The other OGs included 2 proteins related to the nuclear pore (E3 SUMO-protein Ligase RanBP2, and Exportin 1), the piRNA key-factor Piwi, the Cyclin-dependent kinases 1 and 2 (collected into a single homology group), and SUMO-activating enzyme subunit 2 (involved in protein-sumoylation).

The 5 remaining 8-species OGs represented noisy large clusters of homology where only a few sequences were actually upregulated in germline-related samples (less than 1/3 of the CDSs included in each OG). OrthoFinder homology inference is, indeed, prone to collapsing within the same OG different genes belonging to the same gene family, or that simply share some specific domains. This happens especially when domains are common, in multiple copies within the same proteins, and follow complex pattern of acquisition/loss in the proteome, reflecting a network-like homology of conserved protein regions. An example of this was the collapsing of different Tudor domain-containing proteins within the same OG in the Chapter B analysis: TDRD1, TDRD2, TDRD4, TDRD5, TDRD6, TDRD7, TDRD15, and AKAP1, all ended up within the same cluster because the software could not resolve the tangled evolution of the relatively recently evolved and shuffled Tudor domains within these proteins (see Chapter B).

For instance, the OG0000003 (following OrthoFinder default cardinal nomenclature) included 375 CDSs, but only 39 were germline-related. This OG included exclusively genes that encoded for proteases, but a clear whole-length homology could not be retrieved for the germline-related subset (it is however a signal itself the fact that representatives of such OG were consistently upregulated). We could observe similar cases for noisy OGs that included transcripts coding for sodium-dependent transporters, Kinesin-like proteins, peptidyl-prolyl cis-trans isomerases, and beta-1,3-galactosyltransferases.



	Brachionus manjavacas	Caenorhabditis elegans	Clytia hemisphaerica	Danio rerio	Drosophila melanogaster	Ephydatia fluviatilis	Haliotis rufescens	Ruditapes philippinarum	Schmidtea mediterranea	Xenopus tropicalis		Hsa	Cel	Dme
OG0000666	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	Importin-alpha	kpna2	ima3	pen
OG0001682	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**dCMP Deaminase**	dctc	ZK643.2	CG6951
OG0002149	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**Nuclear Autoantigenic Sperm Protein**	nasp	nasp1	CG8223
OG0004330	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**DNA replication nuclease/helicase 2**	dna2	dna2	CG2990
OG0000003	🟦	🟥	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	*(Proteases)*			
OG0000018	🟦	🟦	🟥	🟦	🟦	🟥	🟦	🟦	🟦	🟦	*(Sodium-dependent Transporters)*			
OG0000221	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟦	🟦	🟦	*(Kinesin-like Proteins)*			
OG0000307	🟦	🟦	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	*(Peptidyl-prolyl cis-trans isomerases)*			
OG0000313	🟦	🟥	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	*(Beta-1,3-Galactosyltransferases)*			
OG0000490	🟥	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	E3 SUMO-Protein Ligase RanBP2	ranbp2	ranbp2	nup358
OG0000771	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	Cyclin-Dependent Kinases 1/2	cdk1/2	cdk1/2	cdk1/2
OG0001043	🟥	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	Piwi	Piwil3	prg1	piwi
OG0002814	🟥	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	**DNA Mismatch Repair Protein Msh2**	msh2	msh2	spel1
OG0003912	🟥	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	**Minichromosome Maintenance 10**	mcm10	mcm10	mcm10
OG0004209	🟥	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**Exonuclease 1**	exo1	exo1	tos
OG0004504	🟥	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	**Deoxyuridine Triphosphatase**	DUT	dut1	dUTPase
OG0004523	🟥	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	**Histone Chaperone Asf1b**	asf1b	asfl1	asf1
OG0005801	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	SUMO-Activating Enzyme Subunit 2	uba2	uba2	uba2
OG0006584	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**DNA Replication Licensing Factor MCM4**	mcm4	mcm4	dpa
OG0006764	🟥	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	**Structure Specific Recognition Protein 1**	ssrp1	hmg3	ssrp
OG0007384	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	Exportin 2	cse1l	xpo2	cse1

Figure 2. Upregulated germline-related OrthoGroups (OGs) shared by 8 species or more. The coloured table represent presence (blue) or absence (red) in different species (columns) of germline-related differentially transcribed genes belonging to different OGs (rows; OGs names on the left correspond to the default cardinal nomenclature by OrthoFinder). No germline-related OGs were upregulated in all 10 species. On the right of each row is reported the annotation of proteins encoded by genes included in the respective OGs: bold names represent proteins associated to DNA-related activities; italicized names included between parentheses represent protein functional classes, since the corresponding OGs included large clusters of homology, and not defined orthology groups. On the left, a table summarizes the gene nomenclature in three model species (Hsa: *Homo sapiens*; Cel: *Caenorhabditis elegans*; Dme: *Drosophila melanogaster*).

We also looked at GMP-associated genes that were previously annotated as expressed in germline/multipotent cell lineages (see Introduction). These genes were namely *piwi*, *vasa*, *boule*, *nanos*, *pumilio*, *bruno*, and *tudor* (referring to *D. melanogaster* nomenclature), and we identified their belonging OGs based on the *D. melanogaster* sequences. Of all seven genes, none were included in ubiquitously shared germline OGs (Figure 3). With the exclusion of *piwi* (already cited before since

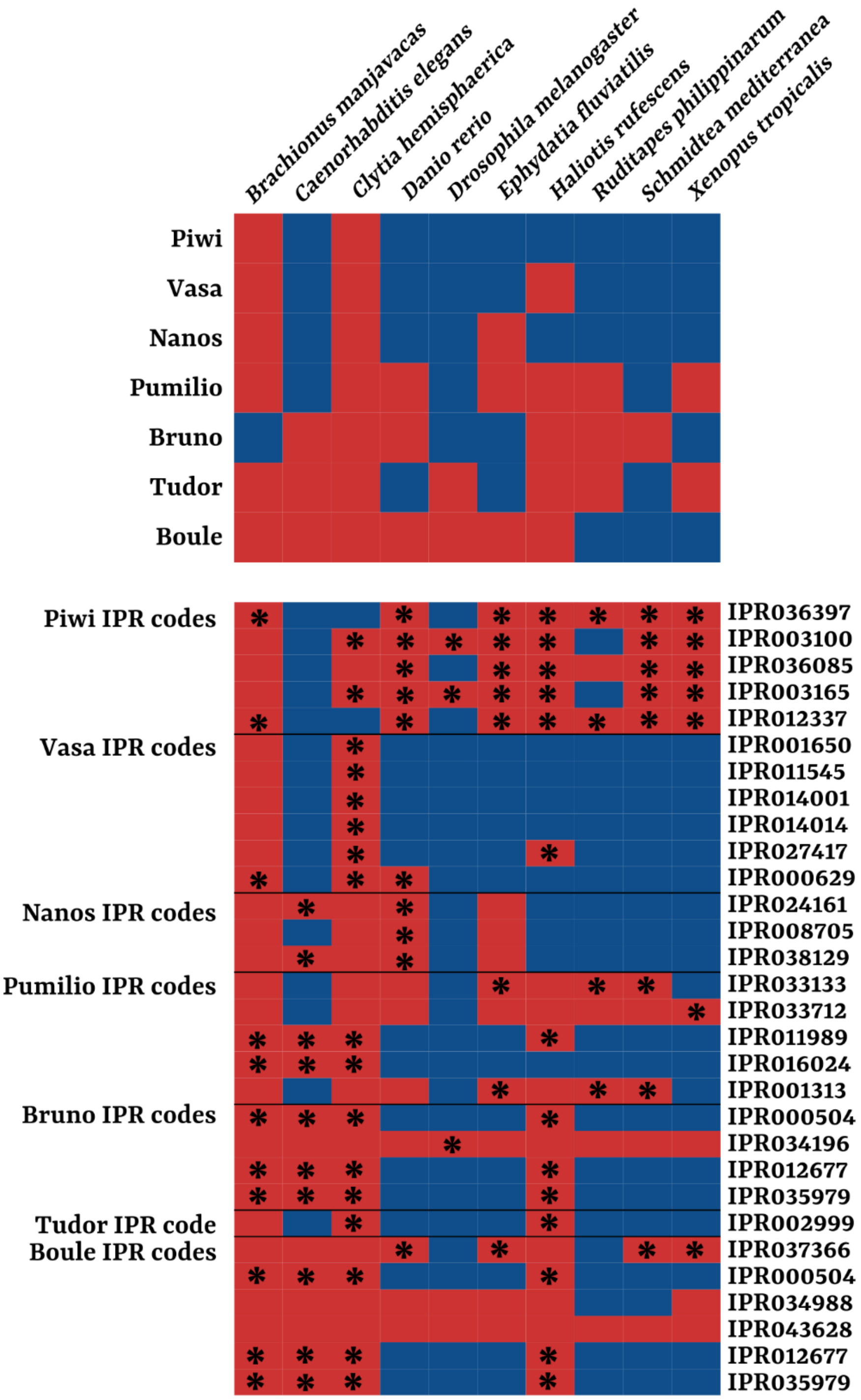


Figure 3. Germline Multipotency Program (GMP) genes differential transcription in germline-related samples. The table on top represent presence (blue) or absence (red) in different species (columns) of germline-related differentially transcribed GMP genes (rows; names refer to *D. melanogaster* nomenclature). *Piwi* is upregulated in germline-related samples of 8 species (present also in Figure 2), while *vasa* and *nanos* in 7 species. Other GMP genes display a more scattered distribution. The table on bottom represent typical GMP domain distribution (rows; names on the left, IPR codes on the right) in each species (columns; for species names refer to the upper table): blue means biased presence of the domain in that species germline-related samples; red means absence of that domain in the germline-related samples; red with asterisks means presence of the domain in germline-related samples, but not significantly biased toward them. If considering the characteristic domains instead of the whole length proteins, the biased presence in germline-related samples slightly rises.

it was included in the OGs shared by 8 species) the only GMP genes that were upregulated in a significant number of species were *vasa* and *nanos* (shared by 7 species). The other ones were shared by only 3 to 4 species. The situation slightly improved when considering IPR codes associated to the proteins instead of the full-length homologous sequences (lower box of Figure 3). For instance, while homologues of *Drosophila tudor* where upregulated in 3 species, the Tudor domain (IPR002999) was biased in the germline-related samples of 7 species.

We then looked at all other OGs that contained sequences upregulated in at least 2 species. The 2-species combinations (i.e. OGs upregulated in 2 species only) were the predominant ones, significantly deviating from the expected random distribution: they corresponded to 2118 of the 4365 germline-related OGs (Figure 4). Of these 2-species combinations, the 4 that displayed a higher degree of positive deviation were the couples *Danio*-*Xenopus* (Chordata), *Brachionus*-*Ruditapes* (Lophotrochozoa), *Clytia*-*Ephydatia* (the only two species basal to Bilateria), *Drosophila*-*Xenopus*, and *Haliotis*-*Ruditapes* (Mollusca), therefore reflecting a weak phylogenetic signal. Interestingly, out of all the combinations of 3 or more species, those that displayed a positive deviation from expected values were 6-, 7-, 8-, and 9-species combinations, while the 3-, 4-, and 5-species ones had negative deviations, hence they were represented in lower numbers in respect to random distributions.

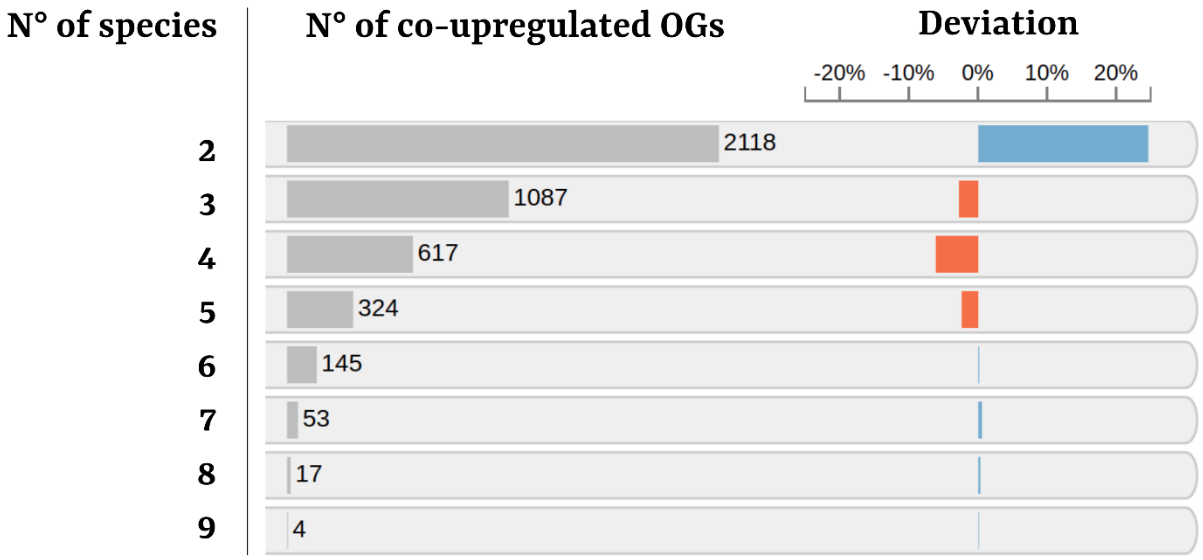


Figure 4. Counts of co-upregulated OGs for all combinations of species. Each row represents the number of OGs that included upregulated germline-related transcripts in a precise number of species (from 2 to 9). For example, first row: 2118 OGs included germline-related upregulated sequences belonging to 2 species only (counting any possible 2 species combination). On the right the deviation from expected random distributions for the combinations of the corresponding number of species is reported: positive deviation from expectation is depicted in blue, negative deviation in red. For instance: the number of observed co-upregulated OGs in 4 species (any 4 species and only 4 species) was lower than expected; the number of observed co-upregulated OGs in 8 species (any 8 species and only 8 species) was higher than expected.

DNA-related functions and domain-specific biases in germline-related samples

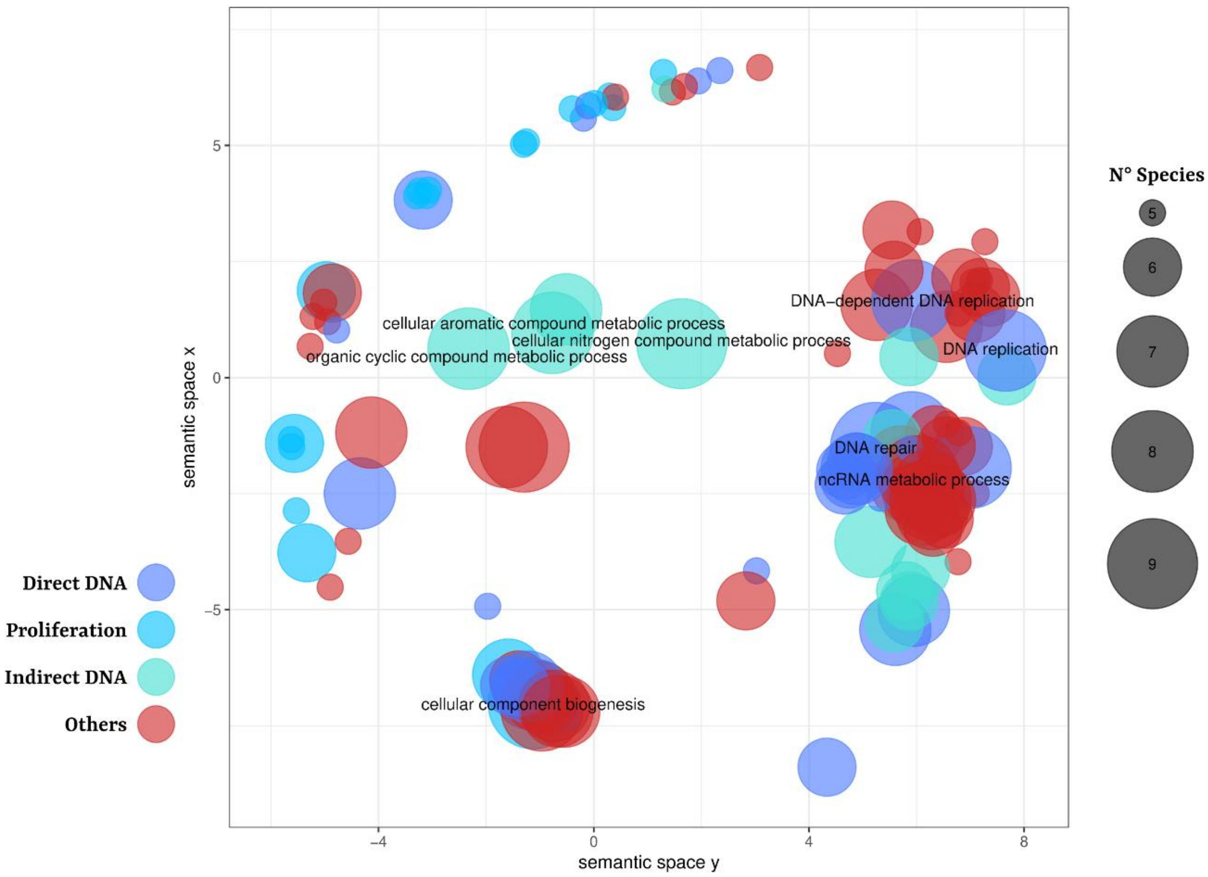
The GO enrichment analysis revealed the presence of many GO terms significantly associated to germline-related samples in 5 or more species of our data set. These results are shown in Figure 5, split in Biological Processes and Molecular Functions (extended results in Supplementary Tables A1 and A2). No GO term was significantly enriched in all 10 species, but 7 were enriched in 9 out of 10 (always excluding *B. manjavacas*, with the exception of one term that was not enriched in *C. elegans*). These included the 2 molecular functions of "Nuclease activity" and "Exonuclease activity", and the 5 biological processes of "Cellular component biogenesis", "Non-coding RNA metabolic process", "Cellular nitrogen compound metabolic process", "Cellular response to stress", and "DNA repair". They all represent generic biological signals, but that can be collectively tied, together with most other GO terms consistently enriched also in less than 9 species, by DNA-related activity in an anabolism-oriented scenario, therefore coherent with previous homology-related results toward DNA replication activity. Indeed, out of 145 GO terms consistently enriched in at least 5 species, 86 were either directly associated to DNA anabolism (e.g. "DNA replication" in 8 species, or "Deoxyribonucleotide biosynthetic process" in 7 species), or more generically associated to cellular proliferation signals (e.g. "Mitotic sister chromatid segregation" or "Chromosome organization", both in 7 species). Also as regards molecular functions, out of 50 enriched GO terms shared by 5 or more species, 18 were directly associated to DNA (such as "Nucleotidyltransferase activity" in 8 species, or "DNA helicase activity" in 7 species).

Signals of strong biases toward DNA-related activity were retrieved also with IPR codes (Table 2). 151 codes were over-represented in germline-related transcripts for more than 5 species. Among these, 112 were represented by codes of domains or families involved in DNA-related activities (from DNA binding to histone regulation and mitotic/meiotic functions), of which more than half (65) were directly associated to DNA replication, and nearly one sixth (17) were related to assessed functions in DNA repair. This signal grew stronger when considering the most shared IPR codes, i.e. those shared by 8 species (again, always lacking *B. manjavacas*, but variable for the remaining missing species): out of 17 such codes, 13 were related do DNA activities, of which 10 directly involved DNA replication (with an almost exclusive representativeness of the mini-chromosome maintenance complex). The remaining 4 codes were associated to RNA helicases, and they were all related to functional domains included in the germline-related helicase Vasa.

Table 2. Biased germline-related IPR codes shared by 7 or 8 species. Codes that are related to direct DNA activity are in bold, and coloured in red when directly associated to DNA replication. “N°” column refers to the number of species for which that IPR code is biased toward germline-related samples.

IPR code	N°	Description	IPR code	N°	Description
IPR001208	**8**	**MCM domain**	**IPR032642**	**7**	**DNA mismatch repair protein Msh2 family**
IPR008047	**8**	**Mini-chromosome maintenance complex protein 4 family**	**IPR033809**	**7**	**USP39 family**
IPR014808	**8**	**DNA replication factor Dna2, N-terminal domain**	**IPR037315**	**7**	**Exonuclease-1, H3TH domain**
IPR015411	**8**	**Replication factor Mcm10, C-terminal domain**	**IPR038167**	**7**	**SSRP1 domain superfamily**
IPR018525	**8**	**Mini-chromosome maintenance conserved site**	**IPR001005**	**7**	**SANT/Myb domain**
IPR027925	**8**	**MCM N-terminal domain**	**IPR001214**	**7**	**SET domain**
IPR031327	**8**	**Mini-chromosome maintenance protein family**	**IPR003593**	**7**	**AAA+ ATPase domain**
IPR033762	**8**	**MCM OB domain**	**IPR003959**	**7**	**ATPase, AAA-type, core domain**
IPR040184	**8**	**Minichromosome maintenance protein 10 family**	**IPR012340**	**7**	**Nucleic acid-binding OB-fold superfamily**
IPR041562	**8**	**MCM, AAA-lid domain**	**IPR019775**	**7**	**WD40 repeat, conserved site**
IPR035417	**8**	**FACT complex subunit POB3-like N-terminal PH domain**	**IPR020472**	**7**	**G-protein beta WD-40 repeat**
IPR015943	**8**	**WD40/YVTN repeat-like-containing domain superfamily**	**IPR027417**	**7**	**P-loop containing nucleoside triphosphate hydrolase superfamily**
IPR029063	**8**	**S-adenosyl-L-methionine-dependent methyltransferase**	**IPR033925**	**7**	**Rad51/DMC1/RadA domain**
IPR014001	8	Helicase superfamily 1/2, ATP-binding domain	**IPR036322**	**7**	**WD40-repeat-containing domain superfamily**
IPR001650	8	Helicase, C-terminal domain	**IPR001680**	**7**	**WD40 repeat**
IPR011545	8	DEAD/DEAH box helicase domain	IPR000571	7	Zinc finger, CCCH-type domain
IPR014014	8	RNA helicase, DEAD-box type, Q motif domain	IPR000629	7	ATP-dependent RNA helicase DEAD-box, conserved site
IPR000969	**7**	**Structure-specific recognition protein family**	IPR002999	7	Tudor domain
IPR001352	**7**	**Ribonuclease HII/HIII family**	IPR007146	7	Sas10/Utp3/C1D family
IPR008048	**7**	**DNA replication licensing factor Mcm5 family**	IPR013026	7	Tetratricopeptide repeat-containing domain
IPR008921	**7**	**DNA polymerase III, clamp loader complex, gamma/delta/delta subunit**	IPR016024	7	Armadillo-type fold superfamily
IPR015408	**7**	**Zinc finger, Mcm10/DnaG-type domain**	IPR017986	7	WD40-repeat-containing domain
IPR016467	**7**	**DNA recombination and repair protein RecA-like family**	IPR024567	7	Ribonuclease HII/HIII domain
IPR024954	**7**	**SSRP1 domain**	IPR028077	7	Ubiquitin/SUMO-activating enzyme ubiquitin-like domain
IPR026851	**7**	**DNA replication ATP-dependent helicase/nuclease Dna2 family**	IPR029060	7	PIN-like domain superfamily
IPR032641	**7**	**Exonuclease 1 family**			

Biological Processes



Molecular Functions

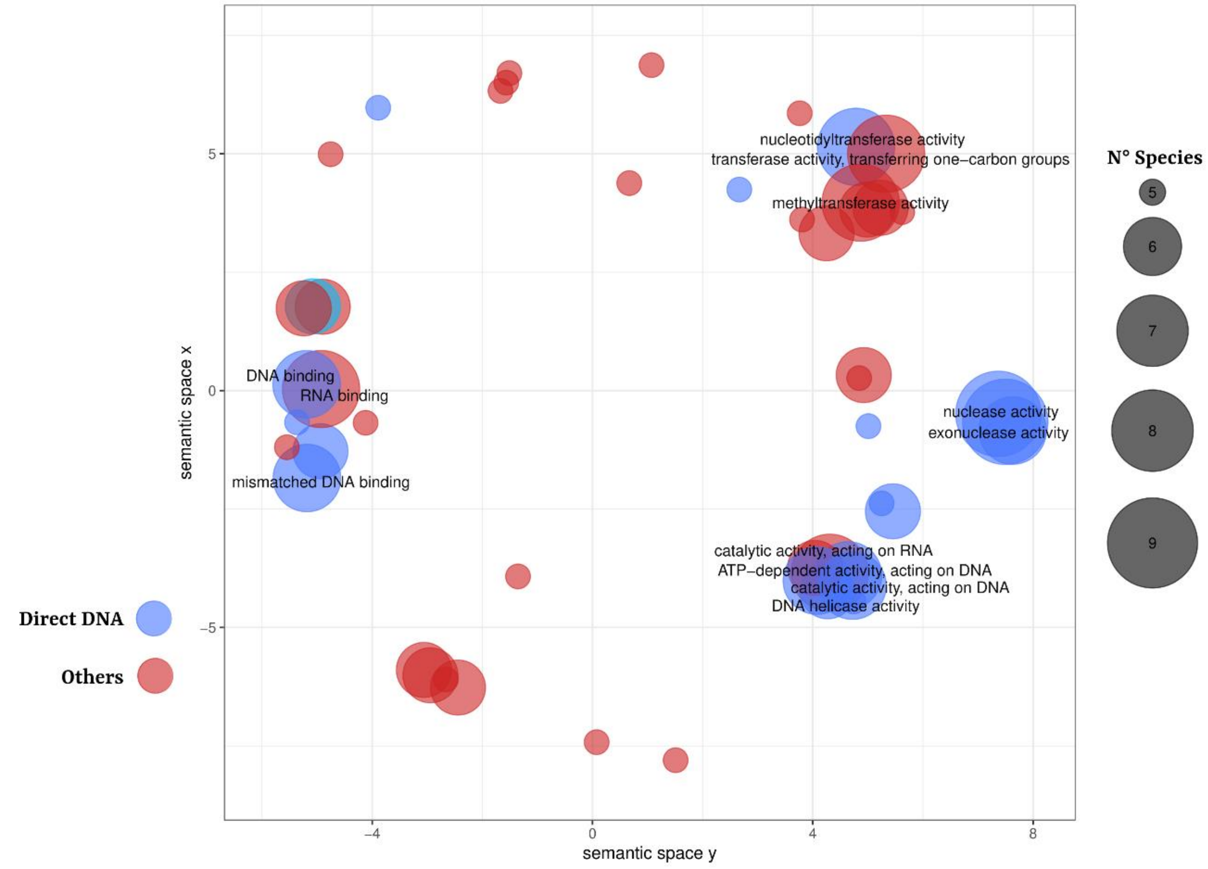


Figure 5. Co-enriched germline-related GO terms shared by 5 or more species. (Figure in previous page) The semantic plot on the top corresponds to GO terms that define biological processes. The semantic plot on the bottom corresponds to GO terms that define molecular functions. GO terms explicated in the figure are those present in more than 7 species with a dispensability lower than 0.5 (a ReviGO value based on distance to semantically close terms; for the full set of GO terms, see Supplementary Tables A1 and A2). In each plot DNA-related terms are depicted in blue (or shades of blue; colours legend on the left). The size of the circles scales with the number of species that share that specific term in their germline-related samples (size legend on the right).

Discussion

Considerations about the reliability of the samples and the analytical approach

An immediately observable pattern from the results of consistently co-upregulated transcripts is the lack of *B. manjavacas* in all clusters, except for those constituted by non-defined genes (Figures 2 and 3; but also from biased IPR codes: Table 2). This could be due to either extreme diversification of this species toolkit for germline specification (see the following sub-chapters), or to biases in sample composition. In fact, *B. manjavacas* samples were constituted by eggs against whole bodies (with eggs removed). There is a double level of possible bias generation in such experimental asset. First, eggs represent the end of the spectrum in the pathways of germline specification, them being the final, differentiated gametes. This might have led to the massive presence of transcripts more involved in subsequent embryogenesis rather than strictly germline-associated ones. Supporting this, its species-specific GO enrichment analysis showed biased signals of many GO terms collected together as "Anatomical structure development" (11 terms), or "Regulation of transport" (24 terms), in respect, for instance, to *D. melanogaster* with 34 terms collected in cell cycle regulation, 29 terms of gene expression regulation, and 15 of cellular processes involved in reproduction (individual GO term enrichments in Supplementary Tables A3 to A12). Secondly, the somatic control samples represented by whole bodies might have additionally diluted and led to missed detection of some germline-associated transcripts. Indeed, the expression of some GMP genes, despite being characteristic to germline-related lineages, is not necessarily restricted to them (Wessel 2016). For instance, *piwi* is also transcribed and expressed in neurons of many species (Kim 2019 and references therein), and the piRNA pathway is crucial for neuronal functions, e.g. memory in mammals (Perera et al. 2019; Leighton et al. 2019). Having a control somatic sample represented by whole bodies, therefore encompassing the whole tissue diversity of the organism (comprising maybe also niches of undifferentiated germ cells), might explain the lack also of highly conserved GMP gene transcription in *B. manjavacas* (Figure 3), casting shadows on the reliability of such experiment itself as regards the approaches and aims of present analysis. Supporting this, CDSs belonging to OG0001043 (that included *piwi* homologues) and OG0000643 (that included *vasa* homologues) were upregulated in the control tissues for this species.

Also *C. hemisphaerica* lacked upregulation of all *a priori* analysed GMP genes (Figure 3). This could be due to the fact that its gonad is highly simplified, with germ cells interposed between two single-cell somatic layers (endoderm and ectoderm, the control somatic tissues for this species in the present study). These surrounding cells are tightly associated to germ cells, regulate their cycle, and have neural type morphologies (Deguchi et al. 2011; Artigas et al. 2018). In our analysis, also in this case *piwi* and *vasa* homologues were differentially transcribed in the somatic samples. However, these genes were transcribed also in the germline-related samples, but three times less than controls (for comparison, *B. majavacas* homologues were differentially transcribed in whole bodies more than 10 times more). The peculiar characteristic of the surrounding cells, and their intimate relationship with the developing gametes (that might also include the production of material for later nurturing) probably led to our results. Nevertheless, in *C. hemisphaerica* some domains or families associated to GMP genes were indeed present among germline-related sample upregulated transcripts (even if IPR codes were not biased toward them; Figure 3), and the patterns of other co-upregulated OGs were not like *B. manjavacas*, but rather comparable to other species (Figure 2). For these reasons we believe that the results obtained for such species should not be considered unreliable.
On the other hand, for all other samples, we could assess the shared presence of GMP genes in the subset of germline-related upregulated transcripts (Figure 3). Especially the three more characterized genes, that are *piwi*, *vasa*, and *nanos*, could be found in almost all samples (all other 8 species for the former, in 7 species for the others), comprising non-classical germline lineages like archaeocytes and neoblasts, that are associated to the same genetic programming of germ cells, as said before (Juliano et al. 2010; Solana 2013). The presence of these signature genes adds solidity to our approach, that despite sample heterogeneity was able to retrieve features common across the species considered, and that were in line with previous work. This allowed us to discuss other transcriptional results that did not comprehend *a priori* characterization of known genes.

Germline-related genes are more frequently conserved across Metazoa

An interesting signal that we retrieved by assessing the percentages of clade-specific genes transcribed in our species data set was the fact that, for many species, genes upregulated in germline-related samples were more conserved across Metazoa than expected by chance (Figure 1). On average, ~86% of each species subset of germline-related transcripts had homologues in at least another phylum (average percentage that rises to nearly 92% when excluding the three species that showed the opposite signal: *B. manjavacas*, *C. elegans*, and *S. mediterranea*; see next subchapter). However, this percentage has no meaning if not compared with the percentage of all other non-germline transcripts that are shared with other phyla and that do not represent lineage-specific innovations. Indeed, when comparing the two percentages, it came clear that there was indeed a bias toward Metazoa-shared genes in germline-related samples (Figure 1). This ratio, that we called phylum-specific ratio, was below 1 in many species, suggesting that, for a newly arisen lineage-specific gene, it is less likely to be involved in genetic pathways associated to the germline. Moreover, when considering genes shared by multiple phyla, we could observe a bias in germline-related samples for the upregulation of genes that share homology outside Metaoza, i.e. with at least one of the 8 Holozoa species included in our dataset (the metazoa-specific ratio of Figure 1 is virtually always lower than 1). Again, coherently, when considering only the species of our dataset, the number of co-upregulated OGs comprising 2, 6, 7, 8, and 9 species deviated positively from random distributions, while OGs comprising 3, 4, and 5 species deviated negatively (Figure 4). Most of the shared transcriptional combinations were those shared by 2 species only, that vaguely reflected phylogenetic relationships (among the highest deviations were the couples *Danio*-*Xenopus*, *Ruditapes*-*Brachionus*, *Ruditapes*-*Haliotis*, and *Ephydatia*-*Clytia*). However, excluding combinations of 2 species only, it is interesting to notice that the combinations comprising more species (6 to 9) were more frequent in respect to random expectations than those comprising less species (3 to 5).

The germline, considered in its wide meaning as any cell that can produce a germ cell (therefore including also pluripotent stem cell lineages of some animals that would otherwise end up in classic Weissman's somatic definition; but also cells like those of the mouse inner cell mass, that can produce germ cells despite most of their fates are somatic; see Introduction; Solana 2013), is one of the most shared cell lineage that can be found in animals. Regardless of whether germline establishment was the adaptive driver of multicellularity (Woodland 2016), or if it was simply one of the first evolving lineages, it is undoubted that its presence represents a major phenotypic trait shared by all animals, given that their last common ancestor was most likely an oogamic multicellular organism (King and Rokas 2017). Our results are coherent with germline early origin since we could observe that newly evolved genes were less likely to get included in such lineage, both as regards newly evolved animal genes, and as regards newly evolved phylum-specific ones.

This signal was particularly strong also for the representative of Porifera in our data set, that is *E. fluviatilis*. The germline-related samples considered in this species were archaeocytes, that are sponges totipotent cells involved both in tissue regeneration, and in sexual and asexual reproduction. Indeed, they can produce both gametes (specifically oocytes) and gemmules, i.e. thousands of packed archaeocytes that are released in the environment where they hatch and give rise to new juvenile individuals (Funayama 2013). This species belongs to an early-branching metazoan taxon (whether they represent the earliest branching clade is still a matter of debate; King and Rokas 2017; Laumer et al. 2019) that has been usually associated with some features that evolved precociously in animal evolution. Archaeocytes themselves have been proposed as being very similar to the ancestral type of animal stem cells (Alié et al. 2015). We could observe a low phylum-specific ratio, suggesting that this cell lineage has indeed a transcriptomic profile that involves genes more conserved in Metazoa and that are datable to older evolutionary times. Coherent results were retrieved in a recent work by Sogabe and colleagues (2019): in the species *Amphimedon queenslandica* (a demosponge like *E. fluviatilis*) they analysed transcriptomes of archaeocytes, choanocytes, and pinacocytes (other two lineages that were proposed as cell states similar to early animal cell lineages) and saw that the percentage of upregulated sponge-specific transcripts was much lower in the former ones. Their number were different from ours in absolute values (different species, methods, and tools), but the ratio of that percentage over the sponge-specific percentage of the whole genome as calculated in their work was curiously similar to our results (0.4). Moreover, like in the present study, also in their analysis the percentage of upregulated pre-metazoan genes was indeed higher in that cell lineage. Lastly, they could also observe strong statistical significance when comparing the archaeocyte transcriptomic profile to that of two holozoan: the choanoflagellate *Salpingoeca rosetta* in the colonial stage (but not in sessile or swimming ones), and the ichthyosporean *Creolimax fragrantissima* in the multinucleate stage (but not in the amoeboid one). They interpreted all these data as the fact that the ancestral metazoan cell type resembled modern transdifferentiating stem cells. What we could observe in the present analysis was that similar inter-phyletic phylostratigraphic signal was actually shared by most of the germline-related samples of the considered species, therefore suggesting us to potentially extend such considerations to totipotent lineages as a whole, and further providing hints on the similarities between stem and germ cell lineages. Remarkably, this signal was shared by nearly the totality of the data set when considering homology outside Metazoa, therefore highlighting how pre-metazoan gene are more likely involved in germline-related pathways than expected by chance. Despite the often mentioned heterogeneity of the data set, both in samples (being sometimes whole gonads, sometimes cell lineages) and stages (being sometimes early stages of differentiation, sometimes late ones), the same signal was obtained for different species despite their supposedly ancestral or derived states (from cnidarian to molluscs and chordates). Unfortunately, availability of experiments suited for our intended analyses were few, but we would be eager to extend the pipeline to other species as soon as new data will be available and see whether the trend still stand.

The opposite trend might suggest lineage-specific adaptations

However, when considering phylum-specific genes the described signal was not shared by all the species of our data set, since the exact opposite trend, i.e. for a newly arisen lineage-specific genes, it is more likely to be involved in genetic pathways associated to the germline, was present for the nematode *C. elegans*, the planarian *S. mediterranea*, and the rotifer *B. majavacas* (but see the first Discussion sub-chapter for the reliability of the data regarding the latter). The possibility that this trend could be driven by the fact that these 3 species had samples comprising cell populations rather than tissues is unlikely. First, also the somatic controls in 2 of these species were represented by cell populations (only *B. manjavacas* has a somatic control constituted by whole bodies depleted of eggs; Table 1), therefore any bias toward the relatively small scale of transcription should have been shared by the two conditions, and it should not have straightforwardly led to the observation of lineage-specific genes being more likely included in germline-related cells. Second, these 3 species were not the only ones with such sample asset: indeed, also the aforementioned *E. fluviatilis* and *C. hemisphaerica* included germline-related samples constituted by cell populations (archaeocytes and growing oocytes, respectively; Table 1). Nevertheless, these two species were characterized by low phylum-specific ratios (0.430 and 0.589, respectively).

In *C. elegans* and *S. mediterranea*, high levels of gene loss have been observed (specific studies of such kind in *B. manjavacas* are lacking). *C. elegans* is a well-known representative of Nematoda, and in such phylum extensive gene loss has been documented, together with an enormous occurrence of orphan genes (i.e. genes that do not share homology with any other taxa), that can only partially be explained by Horizontal Gene Transfer (Rodelsperger et al. 2013; Rodelsperger 2017). Whether these orphan genes are the product of actual *de novo* evolution or of simply artefactual lack of homology detection due to higher evolutionary rates in nematodes is not straightforward to assess (Rodelsperger 2017). However, beside the reasons under it, approximately one third of nematode genomes have no homologues outside the phylum (Rodelsperger et al. 2013), and many genes associated to either key metabolic functions (like heme synthesis genes; Rao et al. 2005) or key developmental pathways (like most Hox genes; Aboobaker and Blaxter 2003) are lost in *Caenorhabditis*.

Also a recent genomic survey on *S. mediterranea* revealed a relatively high level of gene loss (Grohme et al. 2018). In the same study, the authors counted 452 highly conserved genes that were lost in such species, compared to the 284 losses in *D. melanogaster*, and 757 in *C. elegans* (confirming what said in the previous paragraph). Curiously, Grohme and colleagues (2018) could also assess that such species lost most of the genes involved in DNA double-strand-break repair. Since planarians have a very high degree of resistance to γ-radiations, that are known to produce such damages to DNA (Wagner et al. 2011), it is plausible that other genes were co-opted by compensatory pathways, or that completely newly evolved genes did; the same could have happened for other key cellular processes, like those involved in germline-related pathways.

In Chapter B the loss of the Piwi pathway of retrotransposon silencing in some animals is discussed, such as in Neodermata (Platyhelminthes). In these species, new Argonaute proteins evolved and are apparently associated to functions similar to the replaced ones (the FLAgos; Skinner et al. 2014; Fontenla et al. 2021). Also in many non-Clade V nematodes, Piwi has been lost, but nevertheless the load of transposable elements is not different from other species (Szitenberg et al. 2016). Sarkies and colleagues (2015) observed that in these nematodes both ancient eukaryotic pathways and completely newly evolved ones were co-opted to account for the loss of Piwi. The three cases observed in our study of higher phylum-specificity for germline-related transcriptomic profiles, in respect to the rest of the transcriptome, might be interpreted in this way, i.e. the recruitment of novel genes also for highly conserved functions. However, additional analyses should be done on the matter.

These three species were those with the highest absolute values of phylum-specificity of the whole transcriptome. Nonetheless, it is necessary to point out that, even if a genome is characterized by high level of lineage-specific innovations, this should not straightforwardly mean that these innovations should be over-represented specifically in germline-related transcripts. For instance, *X. tropicalis* and *S. mediterranea* in our analysis shared similar levels of whole transcriptome phylum-specificity (19.9% and 21.5% of the transcripts did not have homologues outside the belonging phylum, respectively), but the germline-related phylum-specificity was almost three times higher for the latter (10.3% against 29.4%). Indeed, such signal should not be necessarily and exclusively related to a higher abundance of lost genes, but rather to the genetic flexibility of the species, in particular in the evolution of germline-related pathways.

Interestingly, in the aforementioned study by Grohme and colleagues (2018), they found 1165 flatworm specific genes in the *S. mediterranea* genome, 1104 of which were species-specific innovations. They checked for 626 of these genes in previous RNA-Seq works and remarkably found that their expression was much more enriched in neoblasts rather than in their differentiated progeny (see Supplementary Figures in Ghrome et al. 2018). Neoblasts are undoubtedly characterized by the transcription of classic GMP genes like *piwi* and *vasa*, and this was one of the strongest hints for the definition of GMP itself (see Introduction; but also GMP gene expression of the present analysis, in which interestingly also the homologue of *nanos* is upregulated, despite previously associated exclusively to *S. mediterranea* germ cells: Krishna et al. 2019). However, a broad set of other genes are involved in neoblast characterization, and while an *a priori* determined set of genes is usually investigated, phylostratigraphic analyses are rare, if not lacking. Our results might suggest that neoblasts, despite sharing many transcribed genes with germ cells and other totipotent stem cell lineages, might be characterized by a relatively high use of novel genes, therefore making them a more derived cellular lineage than previously thought. Gehrke and Srivastava (2016) compared planarian neoblasts with neoblasts of Acoela, them too having extensive body regeneration capabilities. Acoela are a taxonomic unit with a controversial phylogenetic position, for which evidence so far led mostly toward them placed as an early-branching bilaterian clade, sister to all other Bilateria (Ruiz-Trillo and Paps 2016). Despite many cellular similarities between neoblasts of these two phyla, and the fact that in both cases expression of at least some GMP gene have been observed, it is not clear yet whether the two cell lineages can be actually considered homologues, therefore suggesting them as a cell lineage present in the common ancestor of Bilateria, or evolutionarily convergent (Gehrke and Srivastava 2016). Deep investigations on the molecular pathways involved upstream, within, and downstream these cell lineages are necessary to answer the question. The biased contribution of lineage-specific novelties that we observed in the present study temptingly leads toward a more derived phenotype. However, beside the fact that we lacked data on Acoela to make a useful comparison, proper genetic network investigations are needed.

Consistently upregulated OGs are coherent with germline-related totipotent processes and biased toward proper DNA replication

When looking at homologous genes that are upregulated in the majority of the samples (8-9 species), it is clear how most of them can be collected in DNA-related activities. If considering also nuclear import/export activity the bias grows stronger, arriving to comprise 13 out of 16 OGs that were co-upregulated in 7 species or more (excluding the 5 noisy OGs; see Results). Such signal was also very strong when considering shared significantly enriched GO terms (Figure 5) and biased IPR codes (Table 2). These enriched functions actually represent basic cellular processes that are however associated with proliferation and mitotic/meiotic activity. Usually germline-related genes are associated to RNA-binding activities, therefore mostly to post-transcriptional regulation, including all famous GMP core genes (Cinalli et al. 2008), with the exclusion of Tudor proteins. These components, that are indeed crucial, are associated to a supramolecular common feature of stem/germ cells, that are cytoplasmic granular regions collectively called germ granules (see Introduction and Chapter C). These structures are ribonucleoproteic (RNP) granules that have diverse degrees of complexity but all share the presence of master GMP regulators that are necessary for the maintenance of totipotency, and especially for post-transcriptional regulation (Cinalli et al. 2008; Voronina et al. 2011). The classic GMP genes indeed represent the necessary components for such lineages, acting as determinants and regulators of the totipotent state, but many molecular factors of the actual phenotypes could intuitively be genes associated to proliferative signals and cell cycle progression, like those that we found. For instance, Onal and colleagues (2012) observed that neoblast-associated gene cluster GO terms, beside RNP-mediated post-transcriptional regulation, are enriched for DNA replication and cell cycle regulation, and transcriptional regulation and chromatin organization.

The subset of consistently upregulated OGs in our data set comprehended indeed many genes that encodes for proteins associated to DNA replication rather than other DNA-related activities like transcription (subsequent protein names refer to nomenclature in vertebrates; see Figure 2 for nomenclature in other model organisms). For instance, DCTC (upregulated in 8 species) and DUT (in 7 species) are metabolic enzymes that produce dUMP (from dCMP and dUTP, respectively; Weiner et al. 1993; Mol et al. 1996). This metabolite represents the upstream step of dTMP, a precursor of dTTP, whose metabolic end is represented by the inclusion of a thymine in the DNA molecule. The enrichment of these two genes hints for a bias toward DNA synthesis in respect to RNA synthesis, therefore toward cell replication, and they are essential for proper DNA replication by balancing metabolite composition toward dTTP production and therefore avoiding dUTP mis-incorporation in the DNA molecule (Mol et al. 1996).

Other gene products are directly involved in DNA replication initiation (like SSRP1, MCM4, and MCM10; Meager et al. 2019; Falbo et al. 2020) and DNA replication progression (like DNA2, ASF1B, and NASP; Richardson et al. 2006; Abascal et al. 2013; Thangavel et al. 2015). The over-representation of precisely DNA-replication-associated factors in respect to transcriptional activators and promoters might suggest a higher level of conservation among the species of such key-cellular process. The regulation of transcription might be more lineage-specifically tuned and defined, leading, for instance, to the complete lack of any transcription factor in the set of co-upregulated OGs (Wagner and Lynch 2008; Schmitz et al. 2016; but see also metazoan transcription factor variability in: de Mendoza et al. 2013). Germ cell specification and programming has been usually associated to transcriptional repression rather than activation, especially in the first stages of differentiation (Seydoux and Braun 2006; Cinelli et al. 2008). During the first steps of PGCs specification in the embryo of model organisms, the retention from somatic differentiation have been associated to transcriptional repression either globally, like repression induced by *pgc* in *D. melanogaster*, and *pie-1* in *C. elegans*, or specifically, such as the case of *blimp1* in *Mus musculus* (Nakamura and Seydoux 2008; Robert et al. 2015).

However, once the lineage has been established and germline-specific transcription is activated, the maintenance of germline fate is apparently delegated to other mechanisms, such as chromatin remodelling and, most of all, those based on mRNA processing, i.e. post-transcriptional regulation, including the activity of many GMP genes in perinuclear RNP granules (Nakamura and Seydoux 2008; Cinelli et al. 2008; Robert et al. 2015). Indeed, also previously cited results by Onal and colleagues (2012) in planarian neoblasts revealed similar signals. In our study, the only 4 germline-biased IPR codes shared by 8 species that did not refer to DNA replication or nuclear import corresponded to domains or families typical of RNA helicases involved in mRNA homeostasis, such as the nearly ubiquitous germline marker Vasa, i.e. in post-transcriptional regulation (Liu et al. 2006). However, consistently shared OGs, and more than half of consistently biased IPR codes, included almost exclusively replicative signals instead of those related to transcriptional or post-transcriptional RNA regulators (except for Piwi). While these mechanisms are undoubtedly crucial for germline maintenance, they might be controlled by different and specific factors distinctly tuned in the various organisms. For instance, the aforementioned master transcriptional suppressor *pgc* (upregulated in *D. melanogaster* in the present study) has no orthologues outside *Drosophila*, and also *pie-1* was observed in the present study as upregulated, but only in *C. elegans*, *D. rerio*, and *X. tropicalis*. Indeed, while RNA processing-related IPR codes and GO terms were enriched in many of the germline-related samples of our data set (e.g. “mRNA processing” shared by 7 species, “ribonucleoprotein complex biogenesis” shared by 8 species; see Supplementary Tables A1 and A2), specific homologous proteins were not, suggesting that what is conserved are the mechanisms, or strategies, rather than the factors involved. Thus, what we retrieved was that the mostly shared specific transcripts/genes were almost exclusively involved in DNA replication, while other key mechanisms appeared subjected to a deeper diversification, sometimes species-specific, as also supported by previous studies.

Germline-related cell lineages represent crucial units for the organism evolution, since they are the carriers of the genetic material in the reproductive/regenerative processes. Indeed, direct comparisons between germline and somatic mutation rates in human and mice revealed that for both species the germline had a number of mutations per base pair per mitosis two order of magnitude lower than the somatic lineage, suggesting adaptive mechanisms to lower the mutation load in germ cells (Milholland et al. 2016). Among the co-upregulated genes in most of our species, there were also those coding for MSH2 and EXO1, two proteins involved in DNA-repair, and specifically in DNA mismatch repair (Graham et al. 2018), that accounts for characteristic DNA damages that follow errors in DNA replication (Li 2008). Moreover, among the germline-biased IPR codes shared by more than 50% of the species, many were directly associated to DNA repair (more than one tenth of the total number of IPR codes). For instance, many of those shared by 7 or 6 referred to domains or families belonged to a diverse set of protein involved in different DNA repair strategies: RecA/Rad51 genes, involved in proper resolution of meiotic homologous recombination and stalled replication forks (7 species; Robu et al. 2001; del Val et al. 2019); Msh2-related proteins, involved in DNA mismatch repair (7 species; Kumar et al. 2019); Exo1, involved in a wide variety of repair mechanisms, from double-strand break repair, to telomere homeostasis (7 species; Sertic et al. 2020); XPG-I-associated domains, present in many proteins with functions of nucleotide excision repair mechanisms for error-free DNA replication and a variety of DNA damages (6 species; e.g. FEN1 and XPG: Balakrishnan and Bambara 2013; Tsutakawa et al. 2020); ERCC4-related domains, characteristic of proteins with key roles in nucleotide excision repair mechanisms, double-strand breaks repair, and telomere proper replication (6 species; e.g. the XPF/Rad1/Mus81 protein family; Zhu et al. 2013; DeMuyt et al. 2018; Arora and Corbett 2019; Sabatella et al. 2021).

Among the different biological and evolutionary mechanisms to lower mutational load, a higher percentage of transcripts that encode for DNA repair factors should intuitively be promoted when DNA replication fidelity is important. Interestingly, interspecific comparisons between livers of long-living and short-living animals showed that the transcription of DNA repair-associated genes was significantly higher in long-living species, coupling the transcription level to the efficiency of the mechanism (McRae et al. 2015). The importance of correct transmission of genetic information across generations, that being the result of sexual or asexual reproduction, or in regenerative processes, is probably the driver of the shared upregulated transcription that we observed in germline-related samples of the analysed species.

Conclusions

Considering that our samples consisted of heterogeneous conditions, that all included germ cells (or totipotent germline-associated cells) but that were not standardized at the same differentiation stage, we were not surprised by the lack of quantitatively strong shared signals. However, it was nevertheless possible to assess that the biased transcription in the germline-related samples of virtually all species led toward proliferative activities, especially DNA replication and cell cycle progress, whose correct and proper course is fundamental for the genetic "responsibility" of totipotent lineages. Usual signals of either transcriptional or post-transcriptional regulation were not massively shared, suggesting a more conserved genetic toolkit for proper genetic inheritance transmission. Moreover, interestingly, for many species, the phylostratigraphic analysis revealed that, based on the level of novel gene occurrence in each species, novel lineage-specific genes are less likely to be included among germline-related upregulated transcripts than how it is expected by chance. This was true also for the sponge archaeocyte lineage, confirming their genetic ancestrality, but not for the other totipotent stem cell lineage included, that are planarian neoblasts, suggesting that they indeed include a high amount of genetic innovations and therefore might represent a more derived phenotype than previously thought.

Note:

The results exposed in the present chapter, are currently being processed for submission to a journal with IF.

Chapter B

The Tudor Domain Family: Shaping Factors Involved in the Evolution of Tudor Proteins in Animals

Introduction

The Tudor domain is a protein-protein interaction domain that has been observed in multiple different proteins shared by a diverse set of eukaryotic organisms. It was originally identified as a repeated sequence of approximately 60 amino acids in the *tudor* gene of *D. melanogaster*, that harbours 11 repeats of the domain (Ponting 1997), but in the last two decades it has been characterized in many other proteins that span a relatively wide range of functions. The domain function has been initially described as binding to symmetrically di-methylated arginines (sDMA; Coté and Richard 2005), but lately it has been shown how also methyl-lysines (in mono-, di-, or tri-methylated states) and asymmetrically di-methylated arginines can be its substrate (Chen et al. 2011; Botuyan and Mer 2016). Its functional activity is due to its tridimensional folding into a five-stranded β-barrel that forms an aromatic cage for methylated residues (Selenko et al. 2001), a core structural composition shared by other domains of methyl-lysine readers, namely MBT, PWWP, Agenet, and the Chromo domain. Such similarity led to ascribe all these domains in a remote-homology superfamily shared among Eukaryota, namely the Royal family (Maurer-Stroh et al. 2003), of which however only Tudor includes methyl-arginine binding. Structural similarities to the Tudor domains have been also found in few proteins of Bacteria, like in the cyanobacterial PSHCP protein, or in *E. coli* ProQ (Gonzalez et al. 2017; Bauer et al. 2019), with RNA-binding activities. Authors suggested that this might represent an evidence of homology, proposing nucleic binding activity as the ancestral function of the domain (Bauer et al. 2019), supported by putative additional DNA-binding activity in Tudor domains of eukaryotic ARID4 (Gong et al. 2013) and TP53BP1 (Lancelot et al. 2007). However, the extremely low distribution of Tudor-like domains in Bacteria might also suggest that the similarity could be due to either convergent evolution, or more likely horizontal gene transfer events from eukaryotes (Gonzalez et al. 2017).

Be that as it may, it is undoubted that the Tudor domain is a distinctive characteristic of Eukaryota, where it was optioned in a diverse set of highly specialized proteins and often in combination with multiple other domains (see Table 1). The functions of these Tudor domain-containing proteins (from now on referred to as Tudor proteins) are usually conserved among species and are involved in a great variety of mechanisms, including transcriptional regulation, DNA repair, DNA methylation, heterochromatin formation, protein N-glycosylation, mRNA splicing, miRNA-mediated RNA silencing, and TEs mobility repression through piRNAs (Chen et al. 2011; Pek et al. 2012; Lu and Wang 2013; Botuyan and Mer 2016).

Based on the secondary structure, and precisely based on N-terminal (N-t) structural extensions to the characteristic β-barrel, Tudor domains belonging to different proteins have been ascribed from three to four groups (Jin et al. 2009; Ying and Chen 2012). The first group comprises those Tudor domains that display no further extension to the core of 5 antiparallel β-strands (referred to as G0 in the present study). A wide range of proteins harbours Tudor domains of such kind, and, although they have been mostly described in animals, they can all be found in other lineages of Eukaryota, suggesting that their appearances predated the evolution of most eukaryotic lineages. Most of the proteins that display domains belonging to this group are associated to histone "reading" activities, spacing from gene expression activation (like TDRD3, KDM4, SGF29) and repression (like PHF1, ZGPAT, ARID4A-B), cell cycle regulation through p53 stabilization (like SETDB1 and PHF20), DNA double-strand break repair (TP53BP1), DNA re-methylation following replication (UHFR1) and histone docking platforms for heterochromatin assembly (LBR) (see Table 1 for more precise information and references). However, the core Tudor domain is present also in proteins with non-histonic function, like FMR, that regulates mRNA stability and localization, i.e. post-transcriptional regulation (Ascano et al. 2012).

All these Tudor proteins share the same characteristic secondary structure of the Tudor domains, however, while most of them harbour a single Tudor domain (or at most a couple of unrelated ones), some have two that are closely associated to one another in their tertiary folding, forming the so-called Tandem Tudor domain (namely TP53BP1, SGF29, SETDB1, and UHFR1). In the case of KDM4A-B, on the other hand, the Tandem Tudor domain displays a further degree of intimacy, with an interdigitated structure thanks to two swapped β-strands that produce a long and continuous β-sheet that joins the two Tudors (called Hybrid Tudor domain; Botuyan and Mer 2016). All Tudor domains belonging to these proteins, with the notable exception of TDRD3, are characterized by the specific binding to methylated lysines.

A different secondary structure is present in the Tudor domains belonging to the second group (referred to as G1 in the present study). These domains, in fact, display an additional α-helix to the canonical β-barrel in N-t position. Such domain has been characterized for the closely related proteins SMN1 and SMNDC1, both components of the SMN complex that has multiple functions related to assembly, metabolism, and transport of different ribonucleoproteins, comprising, but not limited to, the assembly of spliceosomal small nuclear ribonocleoproteins (Kolb et al. 2007). The Tudor domains of these proteins are involved in the protein-protein interactions between the SMN complex and other proteins, like the spliceosomal Sm proteins (Chen et al. 2011). Differently from G0 Tudor proteins (but similarly to TDRD3), SMN1 and SMNDC1 bind di-methylated arginines (specifically the symmetrical ones; Chen et al. 2011). The other protein that displays a Tudor domain with a similar structural conformation (G1) is ALG13, that has a completely different function from all other Tudor proteins: it is indeed a UDP-N-acetylglucosamine transferase involved in key processes of protein N-glycosylation (Gao et al. 2005). However, precise substrate binding of its Tudor domain and biochemical data in animals are lacking.

An additional N-t extension can be found in the Tudor domains of SND1 and of the metazoan germline-related Tudor domain-containing proteins, that all display methyl-arginine binding activities (domain group referred to as G2 in the present study). All these proteins harbour domains with two additional β-strands and one α-helix that precede the β-barrel core. Previous authors considered SND1 and the other proteins of this group separately, based on functional reasons (Jin et al. 2009; Ying and Chen 2012). Indeed, SND1 was initially discovered as a transcriptional co-activator promoting EBNA2-dependent transcription, but lately has been associated to many other expression regulation pathways: from splicing (Gao et al. 2012) to RNA interference through RISC-mediated miRNA (Caudy et al. 2003), but also in stress response (e.g. Gao et al. 2010) and other functions (reviewed in Gutierrez-Beltran et al. 2016). On the other hand, the germline-related Tudor proteins (sometimes referred to as the TDRD group: TDRD1-2-3-4-5-6-7-9-10-12-15, and STK31; but also the proteins Vreteno and Krimper of *D. melanogaster*) are strictly associated to the piRNA pathway of retrotransposable element silencing. However, also SND1 is partially involved in such pathway, and binding to Piwi (the key regulator of piRNA activity) has been assessed (Liu et al. 2010; Ku et al. 2016). In the present study, based exclusively on secondary structures, we considered them all as ascribed in the G2 group, since the N-t Tudor extensions were the same.

The previous authors that formalized the Tudor domain structural divisions, further proposed a model for the evolution of the different structures (Jin et al. 2009). In this model:

- The ancestral structure of the Tudor domain is the one belonging to G0, that is comprised of the 5 β-strands only (as suggested by the fact that other domains of the Royal family are similarly organized)
- The ancestor of SMN-related Tudor proteins (G1) subsequently acquired the α-helix extension (ALG13 was not considered by the authors).
- The evolution of the G2 domain architecture was explained by the inclusion of a G1 Tudor domain within an SN domain of SND1. The protein SND1 is, in fact, composed by 4 complete SN domains and a partial SN domain closely associated with a G2 Tudor domain. Following the model of evolution of Jin and colleagues (2009), a G1 Tudor domain inserted in proximate

C-terminal position to the 2 β-strands of the 5th SN domain, generating the composite structure of the SN-Tudor domain of SND1. Then, this Tudor domain was co-opted in other genes together with the newly acquired N-t β-strands.

Consistently, while SND1 has been annotated in virtually all eukaryotic lineages (so far not found exclusively in *Saccharomyces cerevisiae*; Gutierrez-Beltran et al. 2016), the other G2 Tudor proteins have more recent evolutionary histories. They have been indeed found only in Metazoa, suggesting a relatively recent evolution and radiation, that might have involved an initial precursor that co-opted the G2 domain of SND1. The presence of at least some members of the latter in basal lineages such as Porifera made some authors suppose that the evolution of such proteins predated the common ancestors of animals (Fierro-Constaìn et al. 2017).

The independent origin of multicellularity in animals was sided by the evolution and/or the expansion of many gene families involved in tissue specification and cell lineage determinations. The Tudor genes were among these, and their evolution was closely associated to those cell lineages characterized by totipotency and multipotency, like germline and stem cell lineages. Animal-specific G2 Tudor proteins are usually characterized by the presence of multiple Tudor domains in their sequence (for instance the aforementioned *tudor* gene of *D. melanogaster*, whose homologue in *H. sapiens* is *tdrd6*), that however do not display a Tandem Tudor tridimensional structure like some G0 proteins. The binding affinity of their Tudor domains to methylated arginines of Piwi and its close homologues (Ago3 and Aub) associates these proteins directly to the piRNA pathway (Siomi et al. 2010; Pek et al. 2010; Chen et al. 2011). The presence of multiple domains allows them to establish multiple contact with different sDMAs, suggesting that they might participate in the pathway as docking platform for spatial organization of the components. In facts, elements like Piwi and the G2 Tudor proteins are usually associated in defined cytoplasmic granules (for instance nuage or Yb-bodies in *D. melanogaster*; Siomi et al. 2011; Juliano et al. 2011; Lim and Kai 2015), and several G2 Tudor proteins have been associated to direct roles in their assembly (e.g. TDRD5 and TDRD7; Yabuta et al. 2011; Tanaka et al 2011; see Chapter C)

The piRNA biogenesis pathway is an RNA-mediated pathway of TE silencing that is private of animals and that is particularly crucial for the proper formation of the germline and of multi/totipotent somatic stem cell lineages that can be found in some metazoan taxa (like planarian neoblasts, sponges archaeocytes, or cnidarian I-cells; Juliano et al. 2010; Juliano et al. 2011; Alié et al. 2015, Fierro-Constaìn et al. 2017). The observation of shared genetic elements among these cellular lineages, in facts, led to the definition of GMP, that is a genetic toolkit that evolved early in animal evolution and was central for the establishment of the segregation of immortal cell lineages from specialized somatic one (see Chapter A).

In the present study, we ought to investigate the evolution of the Tudor domain within Metazoa. Using proteomic data from online available animal genomes, we intended to test the evolutionary model that describes the step-wise accumulation of N-t structures to the domain core. Moreover, by screening 24 phyla we were able to assess the presence or absence of Tudor proteins in the different metazoan lineages and in their closest holozoan relatives, allowing for considerations about the driving forces that might explain the dynamics of reduction and expansions of this protein family.

Table 1. Nomenclature and function of known Tudor proteins.

Protein name (*H. sapiens*)	**Protein name (*D. melanogaster*)**	**Tudor domain**	**Co-occurrent domains**	**Functions**	**References**
TDRD3	TDRD3	Single G0	UBA	Recognizes methyl-arginines on histones and on the C-terminal domain of RNApol-II; positive regulation of gene expression; Included in stress granules, in association with FMR, probably sharing translational repression functions	Linder et al. 2008; Yuan et al. 2020
PHF1-19-MTF2	Polycomb-like	Single G0	PHD	Stymulates catalytic activity of Polycomb repressive complexes 1 and 2, that are histone silencers involved in transcriptional repression	Dong et al. 2020
PHF20-L	MBD-R2	Double G0		Subunit of lysine acetyltransferase complex that acetylates histone H4 and stabilize the tumour suppression protein p53	Cui et al. 2012
SGF29	SGF29	Tandem G0		Component of the SAGA complex, a positive regulator of gene expression	Bian et al. 2011
UHRF1		Tandem G0	UBL, PHD, SRA, RING	Methyl-histone binding protein that recruits DNMT1 to recently replicated hemi-methylated DNA to facilitate efficient re-methylation; sensor of DNA inter-strand crosslinks	Bostick et al. 2007; Liang et al. 2015
LBR	LBR	Single G0	Tm, RS region, Globular region II	Transmembrane protein of the inner nuclear membrane proposed as chaperone-like docking platform for heterochromatin assembly; also involved in cholesterol biosynthetic pathway	Liokatis et al. 2011; Nikolakaki et al. 2017
TP53BP1		Tandem G0	BRCT	Involved in double-strand DNA break repair through promoting non-homologous end joining and inhibiting homologous recombination DNA repair	Bunting et al. 2010; Callen et al. 2013
SETDB1	Eggless	Tandem G0	MBD, SET	Histone methyltransferase that tri-methylates K9 of histone H3, inducing transcriptional repression; regulator of tumour suppressor protein p53	Ayyanathan et al. 2003; Fei et al. 2015
KDM4A-B-C	KDM4A-B-C	Hybrid tandem G0	JmjC, JmjN, PHD	Histone demethylase activity associated to transcriptional activation	Whetstine et al. 2006; Labbé et al. 2014
ZGPAT	ZGPAT	Single G0	ZnF-CCCH	Transcriptional repressor through recruitment of the nucleosome remodelling and deacetylase complex	Li et al. 2009; Gui et al. 2012
ARID4	Hat-trick	Single G0	RBB1 N-t, Arid/Bright, CHROMO	Gene suppressor and epigenetic regulator	Gong et al. 2021
SMN1-2-DC1	SMN-SPF30	Single G1		Components of the SMN complex of ribonucleoprotein assembly; binds spliceosomal Sm proteins, involved in spliceosomal small nuclear ribonucleoprotein assembly	Kolb et al. 2007; Chen et al. 2011
ALG13	ALG13	Single G1	OTU, Glycosyl transferase 28 C-t	N-acetylglucosamine transferase involved in key processes of protein N-glycosylation	Gao et al. 2005
SND1	Tudor-SN	Single G2	SN	Positive regulator of gene expression; spliceosomal small nuclear ribonucleoprotein assembly; miRNA RISC-mediated RNA interference; component of stress granules; involved in piRNA pathway	Reviewed in Gutierrez-Beltran et al. 2016
TDRD1	CG9684/CG9925	Multiple G2	ZnF-MYND	PiRNA pathway; Ago3/Piwi-binding	Chen et al. 2009; Vagin et al. 2009; Ku and Lin 2014

TDRD2	Papi	Single G2	KH	PiRNA pathway, Ago3/Piwi-binding	Chen et al. 2009; Liu et al. 2010; Ku and Lin 2014
TDRD4	Qin	Multiple G2	ZnF-RING	PiRNA pathway; Aub/Ago3/Piwi-binding	Ku and Lin 2014
TDRD5	Tejas	Single G2	Lotus	PiRNA pathway; Aub-binding	Yabuta et al. 2011; Ku and Lin 2014
TDRD6	Tudor	Multiple G2		PiRNA pathway; Aub/Ago3/Piwi-binding	Chen et al. 2009; Ku and Lin 2014
TDRD7	Tapas	Multiple G2	Lotus	PiRNA pathway; Piwi-binding	Tanaka et al. 2011; Ku and Lin 2014
STK31		Single G2	PK	PiRNA pathway; Piwi-binding	Chen et al. 2009
TDRD9	Spindle-E	Single G2	DEAD/DEADH, HELICc, HA2	PiRNA pathway; Aub-binding	Vagin et al. 2009; Ku and Lin 2014
TDRD10		Single G2	RRM	Unknown	
TDRD12	Yb-SoYb-BoYb	Single G2	DEAD/DEADH	PiRNA pathway; Ago3/Piwi-binding	Ku and Lin 2014
TDRD15		Multiple G2		Unknown	
AKAP1		Single G2	Tm, KH	Regulation of mitochondrial functions; binding of PKA regulatory subunits	Livigni et al. 2006
	Krimper	Single G2	ZnF-CCCH	PiRNA pathway, Ago3-binding	Sato et al. 2015
	Vreteno	Multiple G2		PiRNA pathway	Zamparini et al. 2011

Materials and Methods

Proteomes

The data set was built by scanning the online NCBI genome database (https://www.ncbi.nlm.nih.gov/datasets/) for all available animal phyla. We looked for all available RefSeq genomes for which the annotation of protein-coding genes was available, keeping all species for lowly represented phyla, and selecting only a subset of samples for overrepresented ones (such as Chordata and Arthropoda). Once the species were decided, we directly downloaded the whole set of protein sequences from each genome, i.e. the proteomes. Some proteomes of high phylogenetic relevance (e.g. the only representatives of certain phyla) were not present on NCBI but were retrieved from other online databases (see Supplementary Table B1 for source information). The resulting data set consisted of 111 species covering 21 metazoan and 4 holozoan phyla (Supplementary Table B1). In many of the downloaded proteomes we noticed the presence of exact duplicates that were likely the results of database artefacts. For instance, in the *H. sapiens* proteome (assembly GCF_000001405.39), the gene *exoc2* was present seven times despite being the exact same biological sequence: all seven sequences were the same isoforms composed of the exact same 27 exons in the same positions of the same chromosome. Non-biological redundancy might lead to inaccurate results, and, for such reason, we ran CD-HIT v4.8.1 (Li and Godzik 2006) on all proteomes collapsing all sequences that shared 100% of identity and the exact same length. Lastly, we also cleaned the proteomes from the few cases where pseudogenes were included.

Identification of Tudor proteins and Tudor domain extraction

In order to find Tudor domain-containing proteins we first inferred the homology relationships among all sequences of our proteomes. First, we collapsed the proteomes so that exclusively the longest isoform was kept for each gene: this was performed based on unique gene identifiers associated to each sequence header, and it was possible exclusively for a subset of genomes with proper annotation (approximately 70% of the data set). Homology clusters were built on the collapsed proteomes using the software OrthoFinder v2.3.11 (Emms and Kelly 2019) with the --ultra-sensitive parameter, that represents the highest sensitivity. All sequences that were collected within the same OrthoGroup (OG) were considered as homologous. To identify which OGs were represented by Tudor proteins, we performed domain annotation with InterProScan v5.45.80 (Jones et al. 2014) on the whole proteomes of the species. We then identified all sequences that comprehended at least one Tudor domain with an e-value cut-off of 10^{-5}, and by crossing the results with the homology clusters we identified all OGs that included Tudor proteins (Tudor OGs).

Some Tudor proteins have been known for decades and their presence and function have been deeply investigated in model animals (see Table 1). We used the sequences from the known Tudor proteins of *H. sapiens* and *D. melanogaster* (both species are included in our data set) to identify their belonging OGs. These OGs for which the content could be named will be referred to as “annotated”, while all other “non-annotated” homology clusters of Tudor proteins will be referred to with their cardinal number as automatically assigned by OrthoFinder.

Our dataset was very diverse both in terms of phylogenetic span, and in terms of protein family expansion. Indeed, we had samples from 24 holozoan phyla and we intended to cover the evolutionary history of a domain that pre-dated the evolution of animals and that is known to be present in a wide range of proteins covering different functions. For these reasons, we expected high rates of variability in the Tudor domain sequences of our dataset, and we implemented the domain search in order to avoid any detection bias due to the uneven representativeness of both proteins and species in the databases used by InterProScan to infer the domain annotations. For each Tudor OG, we extracted all Tudor domains inferred by the first round of prediction by InterProScan, we aligned the sequences with MAFFT v7.471 (Katoh and Standley 2013), we built an HMM profile on the alignment, and with HMMER v3.2.1 (Eddy 2011) we ran back the profile on all sequences included in the OG in order to retrieve domains in additional proteins. This procedure was repeated iteratively, adding to the profile all newly detected Tudor domain sequences at each iteration, until no more hits were retrieved. With this procedure we were confident that we could account for inner OG diversification and obtain as many sequences as possible. The HMMER iterations were performed with different thresholds for the alignment profile constructions: namely, we changed the parameter that set the threshold for the percentage of gaps in each position of the alignment for it to be included in the profile. We set this threshold to 30%, 50%, and 70% and lately we kept the results for which the lowest number of iterations were performed, or, in case of tie, the parameter for which more domains were predicted.

The tandem Tudor domains (two closely sided Tudor domains) of SGF29, TP53BP1, and UHFR1 were fused into single sequences with our extraction pipeline. To obtain the two domains separately, we aligned them and used the positions obtained from the *H. sapiens* protein to split them (SGF29 positions: Espinola-Lopez and Tan 2021; TP53BP1 positions: Charier et al. 2004; UHFR1 positions: Kori et al. 2019)

Phylogenetic tree inference

For the phylogenetic inference, given the large number of sequences obtained and the wide phylogenetic, we performed some additional steps with the purpose of cleaning the data set from overly divergent domains or possible mis-annotations, therefore to reduce the alignment noise. First, to reduce complexity, we considered only Tudor domains belonging to annotated OGs or to other OGs that included a sufficient ratio of Tudor-containing proteins (therefore excluding OGs comprising only few species or sequences, and large OGs comprising only few Tudor proteins): OG needed to include at least two species and at least 10 overall sequences of which at least 10% of them had a Tudor domain. When multiple Tudor domains were present in the same protein, we constructed a within-OG Neighbour-Joining (NJ) tree using PAUP (Swofford 2003). When domains in different positions clustered separately in monophyletic groups, we considered them separately for all further analyses; when they were nested within each other, all domains from that protein were treated together.

We excluded all Tudor sequences lower than 45 amino acids, that is 3/4th of the Tudor domain core reference length (60 amino acids). Then, Tudor domains of each OGs were aligned separately through MAFFT v7.471, with the *--globalpair* alignment option, that assumes that homology is shared for the whole length of the region of alignment, and with the *--dash* mode, that includes the matching of sequences to online databases of tertiary structures, whose retrieved positional information are used to refine the alignment (Rozewicki et al. 2019). Then, for each OG, we evaluated the quality of the alignment for each sequence with the Transitive Consistency Score (TCS) performed by tcoffee (Chang et al. 2014), and we excluded all sequences with a TCS lower than 50. After the exclusion of every sequence, we aligned back the remaining domains and performed subsequent iterative evaluations, until no sequences with TCS lower than 50 were present anymore. Once these low-quality sequences were removed, we concatenated all Tudor sequences from all OGs and aligned them with MAFFT (*--globalpair --dash*).

The alignment was trimmed with BMGE v1.12 (-g 0.99 -b 1 -h 0.7; Criscuolo and Gribaldo 2010), and Maximum Likelihood (ML) tree inference was performed with IQ-TREE v1.6.12 (Minh et al. 2020). The model of evolution was inferred by BIC with ModelFinder as implemented in IQ-TREE. Ultrafast bootstraps were used as node supports, with 1000 iterations performed. Given the enormous tree space for such wide alignment, we ran 5 different IQ-TREE searches (each starting from 99 parsimony trees for which the likelihood is evaluated, and then performing 1000 steps of topology refinement through likelihood maximization) and we later performed tree topology tests (Approximately Unbiased test, AU; Shimodaira 2002) as implemented by IQ-TREE to assess whether the different resulting tree likelihoods significantly differed to one another.

Since we were interested also in the temporal evolution of the Tudor domain, we decided to use a homologous domain to root the tree. Indeed, the Tudor domains is part of a superfamily of domains tied together by remote homology and shared by different eukaryotic lineages: Agenet, MBT, PWWP and Chromo domains all belong to this superfamily (see Introduction; Maurer-Stroh et al. 2003). Among these, the PWWP domain shares with many Tudor a secondary structure strictly composed of 5 β-strands (Wu et al. 2011). For this reason, it was chosen as outgroup for rooting the tree, given that a similar structure enhanced the possibility to align them to Tudor domains despite remote homology.

Beside the ML tree inference of all filtered Tudor domain sequences, we also run an independent cluster analysis based on the alignment profiles of each OG. For each separate alignment of the different OG filtered domains (length and iterative TCS cutoffs), we constructed an HMM profile with HMMER. Then, the profiles were compared with pHMM-Tree (Huo et al.2017), that compares profiles producing a distance matrix with the PRC algorithm over which a NJ tree is calculated. In this way, we assume the monophyly of the sequences within each profile (that were retrieved through the sequence homology clustering of OrthoFinder), easing the signal and treating each OG as a separate taxonomic unit, therefore reducing noise and improving the resolution of the relative relationships.

Prediction and evolution of the Tudor domain secondary structures

To explore the evolutionary pathways of the Tudor domains in Metazoa, we also focused on the evolution of the secondary structure. Tudor domains have been divided into 4 functional divisions, that corresponded to 3 different secondary structures, based on N-t extensions to the 5 β-strands of the domain core (see Introduction). To investigate the evolutionary patterns of acquisition/loss of such extensions, we inferred the secondary structures of all Tudor domains of our data set with the SSpro v6.0 predictor as implemented in the suite of SCRATCH-1D v2.0 (Cheng et al. 2005). SSpro assigns positions to a 3-class division: whether belonging to α-helixes, β-strands, or unstructured regions.

Predictions were performed on the Tudor domains and on their flanking regions (60 amino acids before and after the start and stop positions of the domain).

For each Tudor OG, we manually checked the secondary structure predictions and annotated each domain to the three groups as based on the N-t extensions described by Jin and colleagues (2009): no extensions (G0), one α-helix (G1), or two β-strands and one α-helix (G2). We considered all domains in the same OG as belonging to the same group, even when not all sequences displayed the N-t extensions (fact most parsimoniously due to annotation and prediction issues, rather than real sequence-specific loss). Anyway, the vast majority of the domains within each OG displayed the exact same structure.

We then investigated the evolution of the secondary structures on the ML tree topology. We tested different models of character evolution with the *fitDiscrete* function of R geiger package. We first tested for the default models: ER, that assumes equal rates between all character transitions; ARD, that assumes a different rate for each character transition; and SYM, that assumes different rates for different character changes, but with equal rates for forward and reverse transitions.

Then, we also tested alternative models that could reflect plausible evolutionary histories of the secondary structure. First, we tested the model proposed by Jin and colleagues (2009; subsequently reconsidered by Ying and Chen 2012; see Introduction) that hypothesized the step-wise accumulation of secondary structures starting from an ancestral G0 Tudor (therefore allowing only for G0→G1 and G1→G2 transitions; the "oneway step-wise accumulation" model). Then we tested the same model with the additional possibility of G2 evolving directly from a G0 structure (the "free oneway acquisition" model). Then we tested a model with any possible transition allowed with different rates but with the constraint of admitting only step-wise acquisitions or losses (therefore excluding direct G0→G2 and G2→G0 transitions; the "bidirectional step-wise" model). Lastly, we tested the possibility of having an ancestral G2 structure and admitting only losses of N-t structures (however admitting G0→G1 putative transitions; the "free oneway loss" model).

Comparisons between models were performed based on the corrected Akaike Information Criterion (AICc), with the best model having the lowest AICc. Ancestral State Reconstruction (ASR) was performed with the *corHMM* function of R corHMM package using the best-fitting model of the aforementioned model comparison analysis.

Statistical analyses

All statistical analyses were performed with R v4.1.2. All analysed distributions were previously tested for normality with a Shapiro test, and the null hypothesis was rejected for them all. Spearman correlation tests were performed with the cor.test function as implemented in R. Correlations were ran between the distributions of Tudor gene numbers for each species (counting base on the annotation upstream to the length and TCS filtering for the phylogenetic analysis, and considering the three structural groups separately) and other distributions of values: the number of *piwi*-like homologues for each species (identified as the OG that included *piwi*, *ago3* and *aub* of *D. melanogaster*, and *piwil1-2-3-4* of *H. sapiens*), the number of proteins for each species included in the *ago*-related OG (identified as the OG containing a*go1* and a*go2* of *D. melanogaster*, and a*go1*, a*go2*, and a*go4* of *H. sapiens*), the number of Piwi domain-containing proteins not included in the two aforementioned OGs, the number of protein-coding genes in each species genome (retrieved as the number of sequences left following isoform collapsing as described for the Tudor OGs), the genome size (retrieved from online databases: NCBI, genomesize.com, and specific databases; see Supplementary Table B1), and the gene density (calculated as the genome size over the number of protein-coding genes). P-values were adjusted using the Bonferroni correction method. Both Tudor, *piwi*-like, and *ago*-related gene numbers did not include isoforms (see the second subchapter of the present section for details). To avoid biases due to the smaller genome sizes and the overall reduced gene content of non-metazoan Holozoa, all statistical analyses were performed on the metazoan dataset.

Results

Distribution of Tudor proteins in homology clusters

We could identify a total amount of 248 OGs including at least one Tudor protein. For 26 of these, it was possible to identify the content, based on the annotation of the model organisms (OGs that will be referred to as "annotated"; see Materials and Methods and Table 1). For the others: some OGs likely represented real groups of homologous sequences, but others comprehended exclusively one species (with different isoforms of the same gene forming the cluster) or included very few sparse Tudor proteins among a high amount of other sequences that did not include Tudor domains. These latter cases might have represented an algorithm construct due to the clustering of domains other than Tudor included in the sequences. All OGs that did not meet our cutoffs (see the subchapter of Materials and Methods describing the phylogenetic inference analyses) were not considered as real Tudor protein homology groups, but nevertheless the Tudor proteins included in them were considered for all numeric statistics. We could identify 12 additional Tudor OGs that did not contain sequences of classic model organisms (resulting in a complete set of 38 Tudor OGs). However, only 4 of these OGs included more than 2 phyla, confirming the annotated OGs as the most widely distributed ones (Figure 1).

We could identify a total amount of 3323 genes containing at least one Tudor domain distributed among the 111 species of the data set. Of these sequences, 178 were ascribed in the 12 additional

Tudor OGs, 279 were either not included in any OG or in excluded ones, while the vast majority (2866) were included in the annotated OGs. Among these, approximately one third of the sequences ended up within the same OG, called OG164 (based on the OrthoFinder cardinal nomenclature). Referring to the annotation of *H. sapiens*, this homology group included the Tudor proteins TDRD1, TDRD2, TDRD4, TDRD5, TDRD6, TDRD7, TDRD15, and AKAP1. In order to split this large OG in the corresponding proteins, whole sequence alignments were not sufficient (given OrthoFinder results) and Tudor domain alignments neither, given the non-linear evolution that followed such domain in these proteins leading to many cases of multiple appearances within the same protein (see Results subchapter describing the ML tree). The simplest way to annotate the sequences was based on the co-occurrence of other domains. In this way we managed to subdivide OG164 in 4 groups (as indicated in Figure 1): a set of sequences that contained exclusively Tudor domains (where *H. sapiens* TDRD6 and TDRD15 ended up), a set including Zinc-finger domains (comprehending TDRD1 and TDRD4), a set including Lotus domains (comprehending TDRD5 and TDRD7), and a set including KH domains (comprehending TDRD2 and AKAP1).

The structural division by N-t extensions of the Tudor domain as inferred by the 3-class division of SSpro allowed us to separate the domains, and consequently the genes that harboured them, in the three groups: G0 (no extensions; 1260 genes), G1 (one α-helix: 313 genes), and G2 (one α-helix and two β-strands: 1750 genes). Our bioinformatic prediction of secondary structures coincided with data from the literature as regards the annotated OGs, therefore supporting the validity of the software predictions and allowing us to consider with confidence the results obtained for the other sequences not included in annotated OGs.

Tudor protein distributions across species and statistical correlations with genomic features

We could individuate the great majority of annotated Tudor genes (the most widespread and studied ones) in almost all animal phyla, including basal lineages such as Porifera and Cnidaria (Figure 1). The total number of Tudor genes included in each species genome, however, differed widely, even if most of the species had a number of genes included approximately between 20 and 50, with a certain amount of variability even within a same phylum (Figures 2 and 3). Interestingly, we could observe that most of G0 and G1 genes were present in most of the unicellular Holozoa basal to animals. However, in these organisms the only G2 gene was SND1, while all other OGs could be found exclusively in animal species (with the notable exception of the ichthyosporean *Ichthyophonus hoferi*).

Considering only Metazoa, an immediate pattern emerged by looking at the lower part of the distribution, that are the 17 species that contained a number of Tudor genes equal or lower than 15 (Figures 2 and 3): 15 of them were endoparasites belonging to 5 different phyla (Platyhelminthes, Nematoda, Orthonectida, Rhombozoa, and Cnidaria), while the other 3 were the urochordate *Oikopleura dioica*, and the placozoan *Thricoplax adherens*. With the exclusion of *Brugia malayi* and *Loa loa*, two nematodes belonging to the Spirurida order, all endoparasites included in our data set

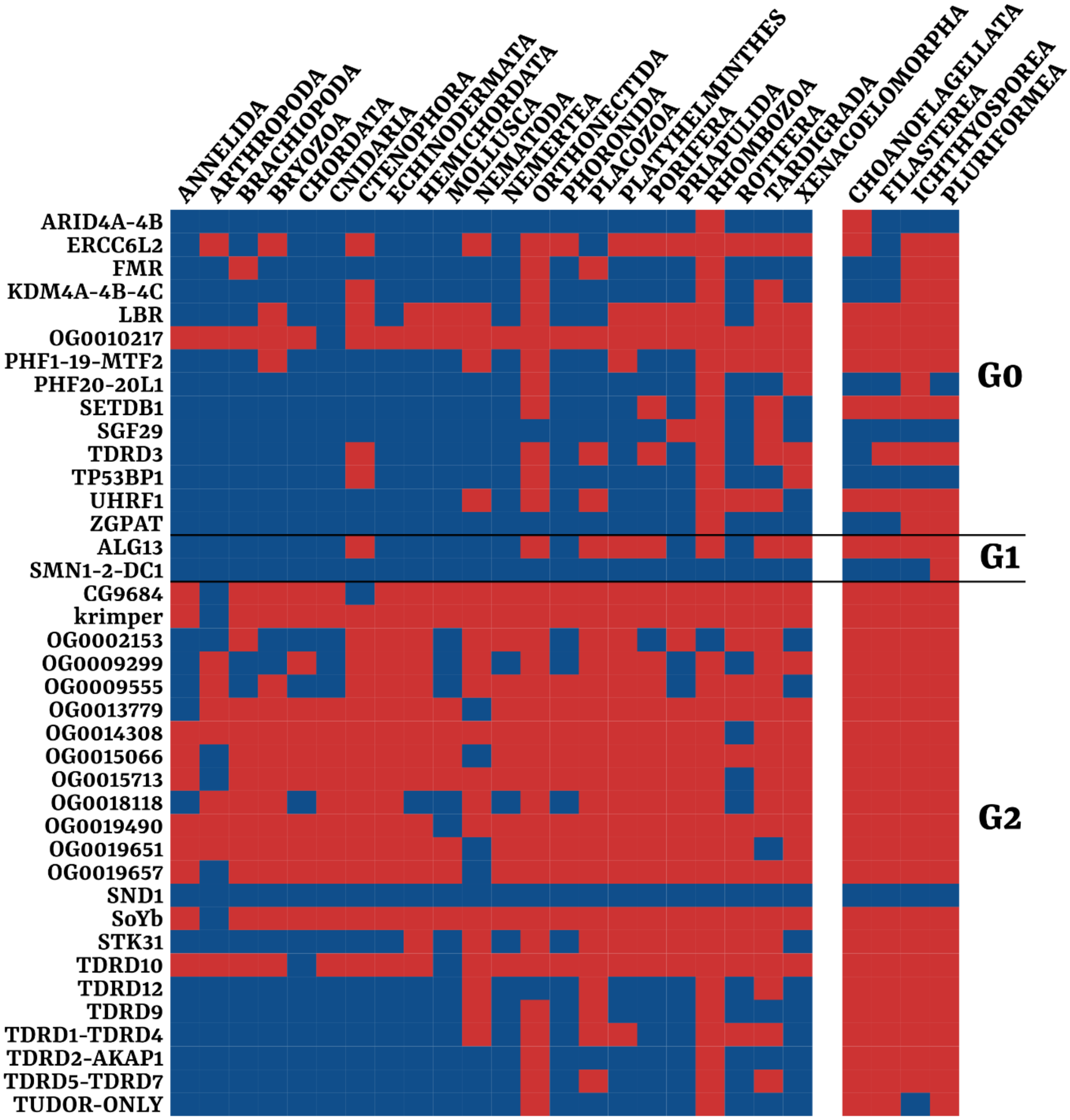


Figure 1. Phylum-specific patterns of presence/absence of Tudor OGs. Tudor OG annotations are depicted on the left side. Known Tudor genes are named with model homologues nomenclature (see Table 1), while others are named with the default cardinal nomenclature by OrthoFinder. The OGs are grouped based on the three possible N-t extensions (group names on the left). The large OG164 is here split in the 4 different components based on the co-occurrence of domains other than Tudor in the protein sequence (see Results). Holozoa phyla basal to Metazoa are separated from animal phyla on the right of the figure. Blue: presence; red: absence.

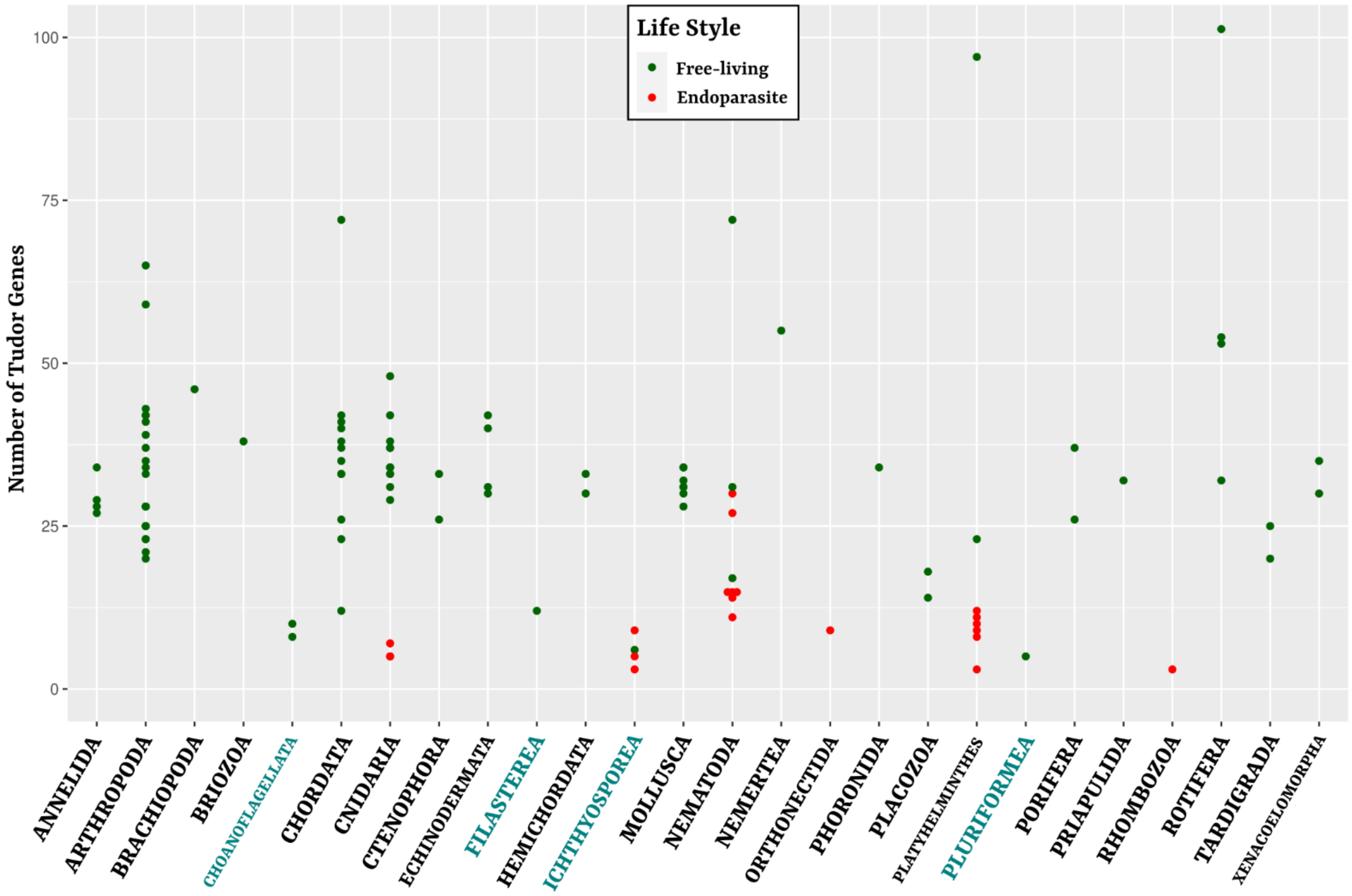


Figure 2. Phylum-specific Tudor gene number distributions. Each dot represents a species, grouped by belonging phylum. Different lifestyle strategies are depicted with different colours (see legend). Non-metazoan Holozoa phyla are highlighted in blue.

ended up in the lower bottom of the distribution, suggesting a shared evolutionary pattern. On the other hand, we could observe some species for which the number of Tudor genes was notably higher than the majority of the species. The species with the highest number of Tudor genes was the bdelloid rotifer *Adineta ricciae* (102 genes), followed by the free-living flatworm *Macrostomum lignano* (97 genes), the sturgeon *Acipenser ruthenus* and the free-living nematode *Plectus sambesii* (both with 72 genes).

We investigated what could be statistical predictors of the observed Tudor gene distribution by correlation tests (all resumed in Figure 4; for species-specific values used in the statistical analyses, refer to Supplementary Table B2). We could assess a significant correlation between the number of Tudor genes in each species and the total number of protein-coding genes in the genome ($\rho = 0.650$; p-value = $5.871*10^{-12}$), together with a significant but weaker correlation with the genome size ($\rho = 0.401$; p-value = $1.456*10^{-3}$). Given that we could find also a weak, but nevertheless barely significant, correlation between the total number of genes and the genome size ($\rho = 0.359$; p-value = $1.057*10^{-2}$), the shape of the Tudor family distribution might have been interpreted just on the light of this shared genomic trend.

However, to further investigate the heterogeneous nature of the Tudor gene set, we divided the distribution into the three components based on the secondary structure: the G0, G1, and G2 groups. When considering them separately, we could observe that, despite the genes belonging to the G2 groups were both more numerous and variable in absolute numbers (Figure 3), the coefficients of variation of the three distributions were similar. Indeed, the G2 coefficient of variation (calculated as the standard deviation over the mean of the distribution) was approximately 63.3%, while for G0 and G1 it was 53.1% and 57.5%, respectively. Nevertheless, these differences in distribution shapes were underlined by differences in statistics correlations. The distributions of all three groups were positively correlated against the total number of genes, but:

- only G0 and G1 were correlated to genome size (G0: $\rho = 0.488$; p-value = $1.752*10^{-7}$; G1: $\rho = 0.363$; p-value = $1.639*10^{-4}$; for G2 there was no significant correlation).
- Both G0 and G1, but not G2, correlated positively and significantly with the genome-specific ratio between the number of genes and the number of proteins (therefore a proxy of the level of alternative splicing; G0: $\rho = 0.438$; p-value = $1.695*10^{-4}$; G1: $\rho = 0.465$; p-value = $3.156*10^{-3}$).
- The G0 distribution was the only one that was negatively correlated also to gene density ($\rho = -0.339$; p-value = $2.502*10^{-2}$).

G2 represents the structural cluster of germline Piwi-associated Tudor proteins. Therefore, we investigated possible correlations between G2 and the number of *piwi* homologues in each species, the number of genes belonging to the Ago proteins subfamily (that share the same Piwi and PAZ domains like *piwi*, indeed belonging to the same Argonaute family), and the number of other Piwi domain-containing proteins that did not belong to homology groups of *piwi* or *ago*.

First, we checked for correlations between these genes and the other genomic statistics that were correlated with the Tudor proteins: we could indeed find a correlation between *piwi* homologue counts and the total number of genes ($\rho = 0.380$; p-value = $4.195*10^{-3}$), but not with any other of the other genomic statistics considered. We observed however correlations between genome size and gene density with the number of Piwi-domain containing proteins not included in *piwi* or *ago*-related OGs ($\rho = -0.394$; p-value = $2.111*10^{-3}$; $\rho = 0.493$; p-value = $6.774*10^{-6}$; respectively). However, *ago*-related genes were not correlated with any of the genomic statistics.

We then looked for correlations against the three Tudor structural groups: *piwi* homologue gene counts were correlated with all groups (as expected since all four correlated against the total number of genes). However, the correlation against G2 was notably stronger (G0: $\rho = 0.478$, p-value = $1.803*10^{-5}$; G1: $\rho = 0.410$, p-value = $9.460*10^{-4}$; G2: $\rho = 0.630$, p-value = $5.549*10^{-11}$). On the other hand, neither the *ago*-related genes nor the remaining Piwi-domain containing genes correlated with any of the Tudor structural group distributions.

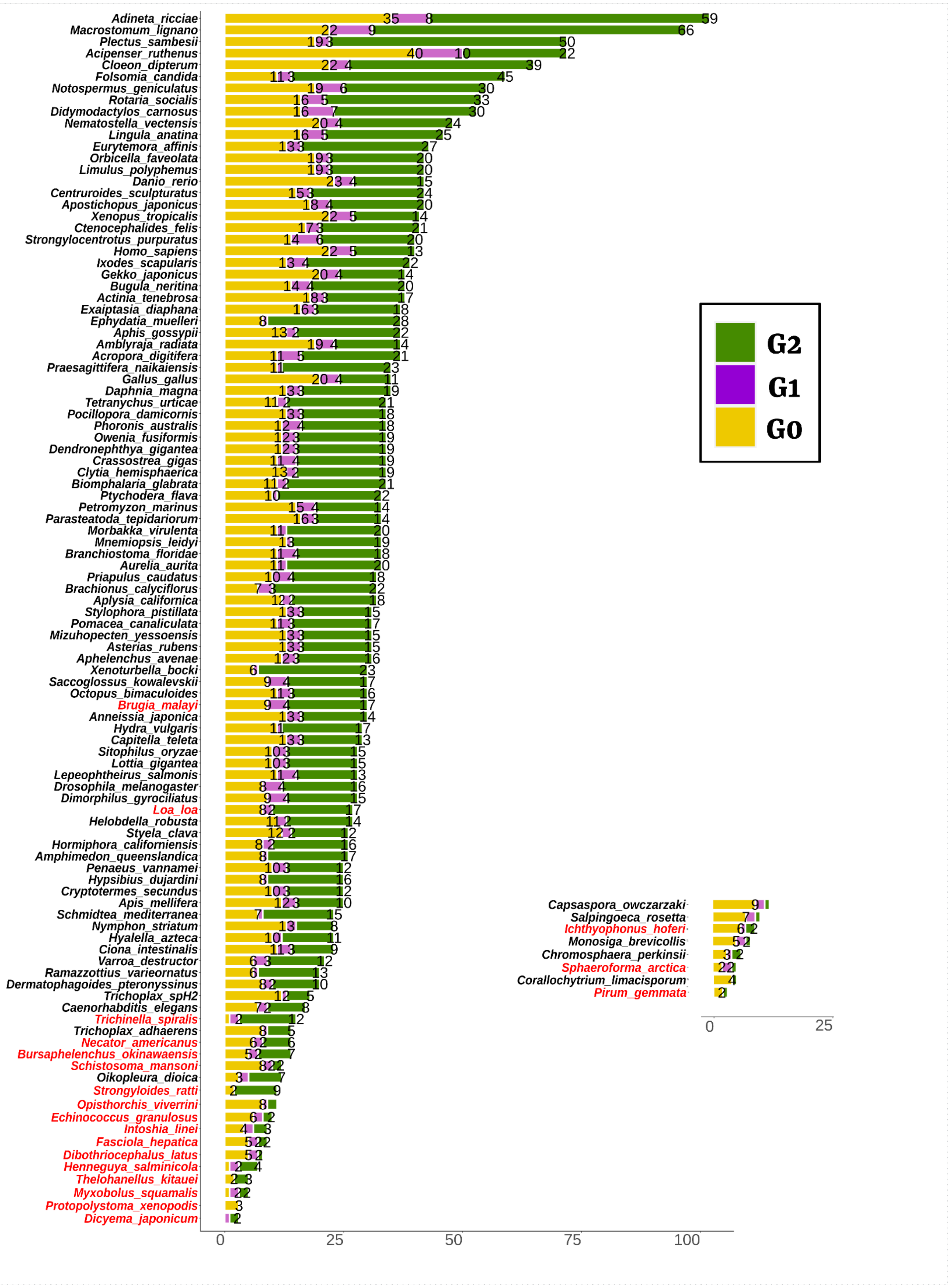

Adineta_ricciae 35 8 59
Macrostomum_lignano 22 9 66
Plectus_sambesii 19 3 50
Acipenser_ruthenus 40 10 22
Cloeon_dipterum 22 4 39
Folsomia_candida 11 3 45
Notospermus_geniculatus 19 6 30
Rotaria_socialis 16 5 33
Didymodactylos_carnosus 16 7 30
Nematostella_vectensis 20 4 24
Lingula_anatina 16 5 25
Eurytemora_affinis 13 3 27
Orbicella_faveolata 19 3 20
Limulus_polyphemus 19 3 20
Danio_rerio 23 4 15
Centruroides_sculpturatus 15 3 24
Apostichopus_japonicus 18 4 20
Xenopus_tropicalis 22 5 14
Ctenocephalides_felis 17 3 21
Strongylocentrotus_purpuratus 14 6 20
Homo_sapiens 22 5 13
Ixodes_scapularis 13 4 22
Gekko_japonicus 20 4 14
Bugula_neritina 14 4 20
Actinia_tenebrosa 18 3 17
Exaiptasia_diaphana 16 3 18
Ephydatia_muelleri 8 28
Aphis_gossypii 13 2 22
Amblyraja_radiata 19 4 14
Acropora_digitifera 11 5 21
Praesagittifera_naikaiensis 11 23
Gallus_gallus 20 4 11
Daphnia_magna 13 3 19
Tetranychus_urticae 11 2 21
Pocillopora_damicornis 13 3 18
Phoronis_australis 12 4 18
Owenia_fusiformis 12 3 19
Dendronephthya_gigantea 12 3 19
Crassostrea_gigas 11 4 19
Clytia_hemisphaerica 13 2 19
Biomphalaria_glabrata 11 2 21
Ptychodera_flava 10 22
Petromyzon_marinus 15 4 14
Parasteatoda_tepidariorum 16 3 14
Morbakka_virulenta 11 20
Mnemiopsis_leidyi 13 19
Branchiostoma_floridae 11 4 18
Aurelia_aurita 11 20
Priapulus_caudatus 10 4 18
Brachionus_calyciflorus 7 3 22
Aplysia_californica 12 2 18
Stylophora_pistillata 13 3 15
Pomacea_canaliculata 11 3 17
Mizuhopecten_yessoensis 13 3 15
Asterias_rubens 13 3 15
Aphelenchus_avenae 12 3 16
Xenoturbella_bocki 6 23
Saccoglossus_kowalevskii 9 4 17
Octopus_bimaculoides 11 3 16
Brugia_malayi 9 4 17
Anneissia_japonica 13 3 14
Hydra_vulgaris 11 17
Capitella_teleta 13 3 13
Sitophilus_oryzae 10 3 15
Lottia_gigantea 10 3 15
Lepeophtheirus_salmonis 11 4 13
Drosophila_melanogaster 8 4 16
Dimorphilus_gyrociliatus 9 4 15
Loa_loa 8 2 17
Helobdella_robusta 11 2 14
Styela_clava 12 2 12
Hormiphora_californiensis 8 2 16
Amphimedon_queenslandica 8 17
Penaeus_vannamei 10 3 12
Hypsibius_dujardini 8 16
Cryptotermes_secundus 10 3 12
Apis_mellifera 12 3 10
Schmidtea_mediterranea 7 15
Nymphon_striatum 13 8
Hyalella_azteca 10 11
Ciona_intestinalis 11 3 9
Varroa_destructor 6 3 12
Ramazzottius_varieornatus 6 13
Dermatophagoides_pteronyssinus 8 2 10
Trichoplax_spH2 12 5
Caenorhabditis_elegans 7 2 8
Trichinella_spiralis 2 12
Trichoplax_adhaerens 8 5
Necator_americanus 6 2 6
Bursaphelenchus_okinawaensis 5 2 7
Schistosoma_mansoni 8 2 2
Oikopleura_dioica 3 7
Strongyloides_ratti 2 9
Opisthorchis_viverrini 8
Echinococcus_granulosus 6 2
Intoshia_linei 4 3
Fasciola_hepatica 5 2 2
Dibothriocephalus_latus 5 2
Henneguya_salminicola 2 4
Thelohanellus_kitauei 2 3
Myxobolus_squamalis 2 2
Protopolystoma_xenopodis 3
Dicyema_japonicum 2
0
25
50
75
100
G2
G1
G0
Capsaspora_owczarzaki 9
Salpingoeca_rosetta 7
Ichthyophonus_hoferi 6 2
Monosiga_brevicollis 5 2
Chromosphaera_perkinsii 3 2
Sphaeroforma_arctica 2 2
Corallochytrium_limacisporum 4
Pirum_gemmata 2
0
25

Figure 3. Species-specific Tudor gene number distributions. (Figure in previous page). Tudor genes are split based on the three structural groups (G0-1-2; see legend for colours). The number of Tudor genes of each group is depicted at the end of each group component of the species bar plots. Non-metazoa Holoza are separated on the right of the plot. Endoparasites are depicted in red.

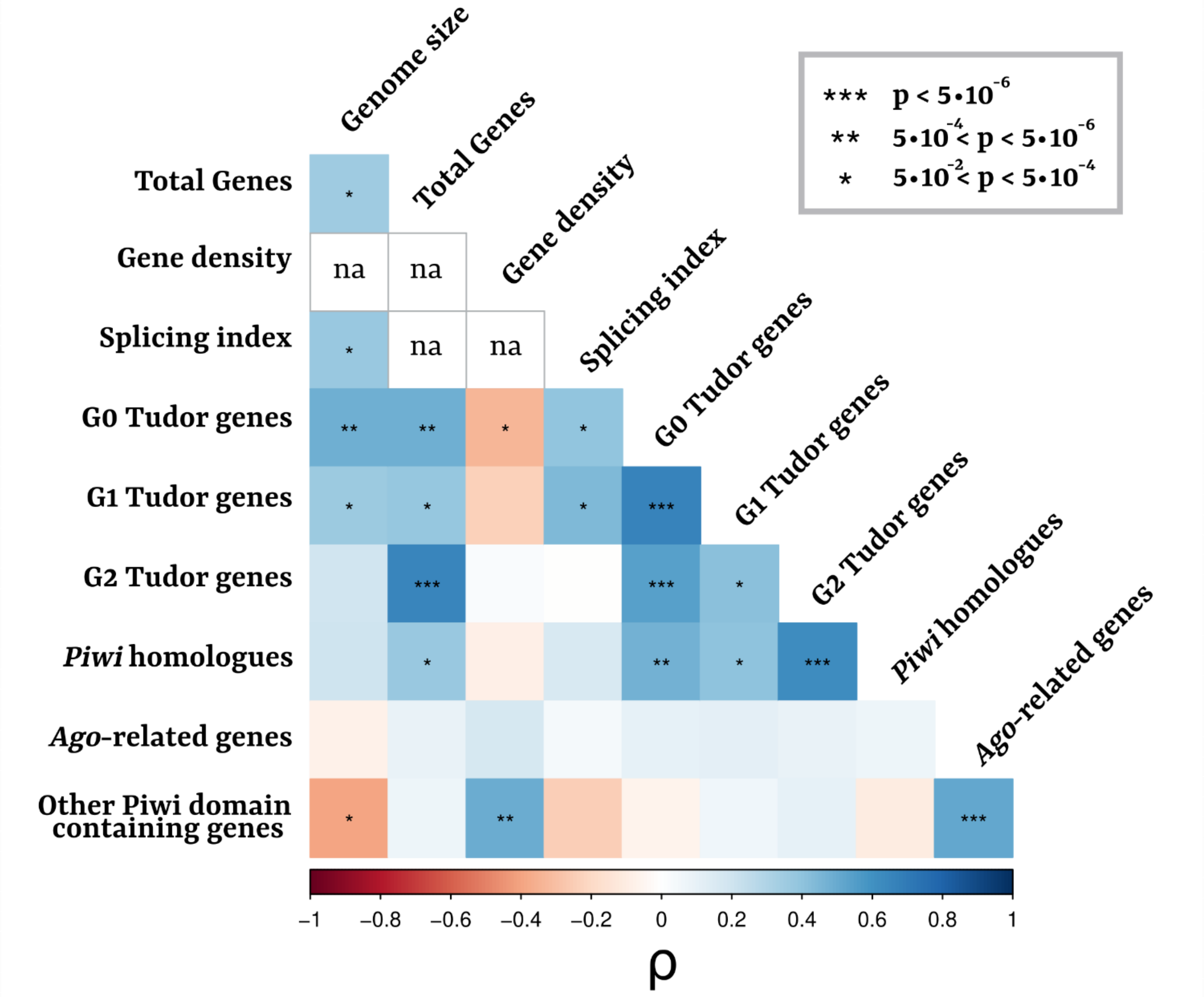


Figure 4. Correlation analyses of Tudor genes with genomic statistics and Piwi domain-containing genes. Tudor genes are split based on the three structural groups (G0-1-2). Correlation coefficient scores are summarized with colours (see bottom legend). Significant correlations are depicted with asterisks (see top-right legend). Some correlations are omitted from the table because the correlating variables were nested among each other, mining their necessary statistical independence.

Phylogenetic inference of the Tudor domain tree

The final set of Tudor domains extracted and aligned for the tree inference included 5158 sequences. To these, 150 sequences of the PWWP domain were added and used to infer the tree root and time-related evolutionary perspective. The best-fitting evolutionary model inferred with the ModelFinder of IQ-TREE for the construction of the tree was LG+F+R10.

The 5 independent run of IQ-TREE inferred trees with likelihood not significantly different from one another (AU topology test). For these reasons, we decided the best one based on biological considerations. For instance, we excluded trees with outgroup paraphyly, or trees with anciently evolved Tudor domains in highly derived positions. The resulting ML tree was highly complex, given the high number of tips, but some considerations could be drawn (Figure 5). Bootstraps support values for each node are depicted as a color gradient: most of the deep nodes were lowly supported, most likely due to the short length of the alignment relative to the number of tips and the evolutionary times considered.

Interestingly, monophyly of Tudor domains of most of the OGs could be retrieved with the ML tree, with the exclusion of OG164 and few sequences that ended up sparsely along the tree. However, a pattern that was common for proteins with multiple Tudor domains, was the separation of the monophyletic groups of the different copies in non-sister relationships. This was the case for instance of SGF29, UHFR1, and SETDB1 proteins: the first and the second Tudor domains present in these proteins clustered separately and in different positions of the tree (for both methods of tree inferences). Only in the profile NJ tree, the two domains of KDM4A-B-C clustered in sister relationship, suggesting a duplication that happened within the protein.

A more complicated pattern was observed for the multiple Tudor domain-containing proteins of the previously cited OG164. Domain sequences from such OG were all included in the clade annotated as B in Figure 5, together with the vast majority of other G2 genes. However, the OG was highly polyphyletic, with many mixed subclades that were related among each other by short branches and lowly supported nodes scattered all along the clade and intercalated by other numerous clades belonging to other OGs. The only clades belonging to OG164 that displayed an apparent "order" were two clades of the Tudor domains of TDRD2 and AKAP1 (both sharing the additional KH domain), that formed separate monophyletic groups in sister relationship between each other. However, for both the TDRD1-TDRD4 subgroup (that include Zn-finger domains), the TDRD5-7 subgroup (including Lotus domains), and the subgroup including only Tudor domains (that included also human TDRD6 and TDRD15), the phylogenetic pattern was too convoluted to confidently retrieve an order. For instance, the three domains of TDRD7 did not form monophyletic clusters but were scattered on the tree.

Not surprisingly, given the massive reduction in complexity, the phylogenetic inference performed through NJ on the distance matrix of the HMM alignment profiles was much clearer (Figure 6). In such tree, all Tudor genes that can be found in most eukaryotes and that share the G0 type of secondary structure represented early branching clades, while all G2 domains were derived, monophyletic and displayed much shorter basal branches. Moreover, also the two G1 domain profiles clustered together.

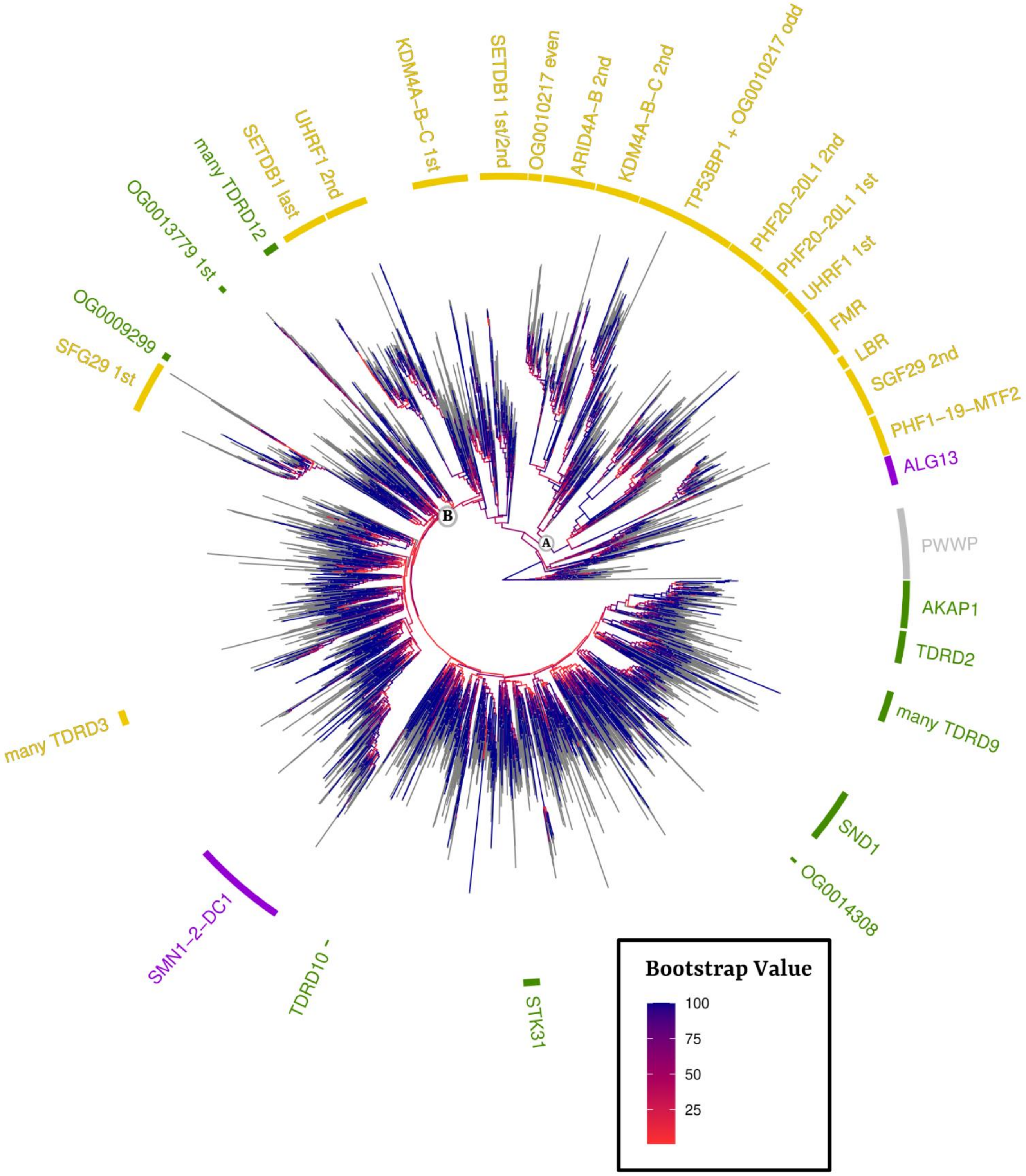


Figure 5. Maximum-Likelihood phylogenetic tree of the Tudor domains. Tree was inferred with IQ-TREE. Here is depicted the best topology out of 5 replicates based on biological expectations (all replicates did not significantly differed from one another in likelihood; AU test) Bootstrap supports are shown as colour gradients (see legend). Monophyletic groups of domains belonging to the same OG were annotated and highlighted with colours based on the structural N-t extensions (yellow: G0; purple: G1; green: G2; grey: outgroup; colours are the same used for Figure 3, 6, and 7).

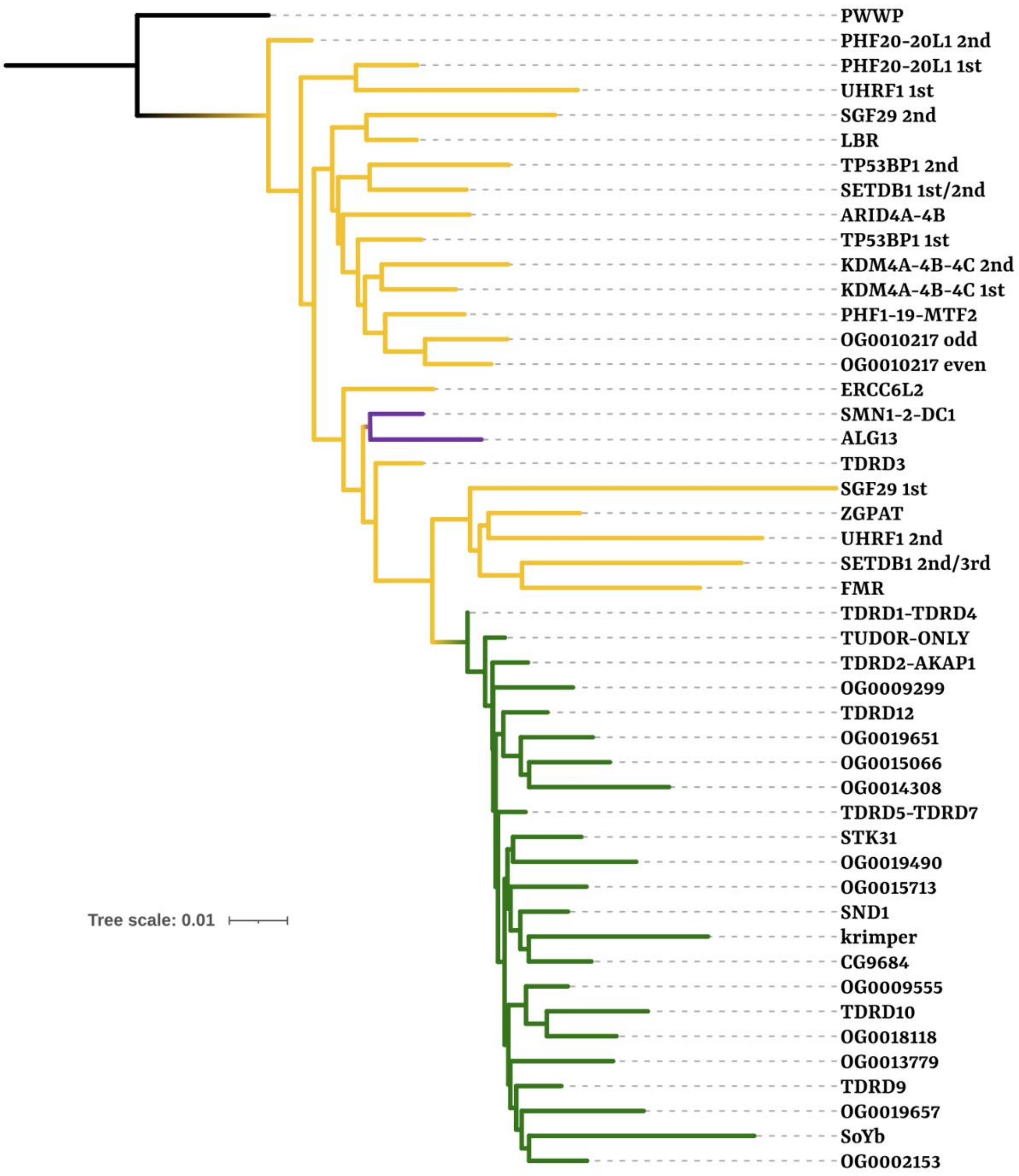


Figure 6. Neighbour-Joining phylogenetic tree of the alignment profiles of the Tudor domains. Tree was inferred with pHMM-Tree. An HMM profile was built for the alignment of Tudor domains belonging to the same OG. The, a distance matrix was calculated between them and a NJ tree was inferred. The tree topology is almost perfectly coherent with expectation from previous hypothesis on Tudor domain evolution Branches are highlighted with colours based on the structural N-t extensions (yellow: G0; purple: G1; green: G2; colours are the same used for Figure 3, 5, and 7).

Evolution of the Tudor domain secondary structures

Some considerations about the evolution of the N-t extension of the Tudor domains can be drawn from a visual inspection of the ML tree. Indeed, almost all early branching clades, after the outgroup separation, were constituted by genes that shared the G0 Tudor domain architecture (i.e. without N-t extensions), and their basal nodes had relatively higher bootstrap support values with respect to most other deep nodes of the tree. The first splitting clade was formed by some sequences belonging to SMN1-2-DC1, TDRD3 and ERCC6L2 Tudor domains, but they were only a limited subsample of them (mostly cnidarian sequences). The first solid clade was formed by all PHF1-19-MTF2 domains in sister relationship with ALG13 ones. Its sister clade was split in two branches: one included a monophyletic group containing domains of most G0 genes (Clade A in Figure 5), the other contained all remaining domains. This latter clade was characterized by subsequently branching clades that comprehended virtually all remaining G0 genes (with the exclusion of the first domain of SGF29 and TDRD3), up to Clade B, which included all G2 domains.

Therefore, the fact that all branches that split before Clade B belonged to G0 Tudors domains while all G2 domains were included in Clade B, support the previous hypothesis on the evolution of the Tudor domain extensions, that sees the acquisition of N-t structures to an ancestral G0-type domain (Jin et al. 2009). Nonetheless, within Clade B, there were some independently nested clades of G0 (first SGF29 Tudor domain and many TDRD3 sequences), that yielded longer stem branches in respect to all other G2 (Figure 5). The HMM profile, on the other hand, yielded an extreme clear topology, with all G2 domains grouped together in a single clade in a derived position (Figure 6)

The two major G1 OGs were not closely related but separated on the ML tree: ALG13 represented an early branching clade close to the tree root, while SMN1-2-DC1 sequences were nested within the large B clade (except for a handful of sequences collected in the small first branching clade of the tree). However, the two clustered together in the HMM profile tree (Figure 6).

To summarize, it appeared that the structural division were good predictors of the phylogenetic relationships between the Tudor domains of the different OGs, and the proposed model of step-wise accumulation of N-t extensions (see Introduction) was partially supported by the mere topology of both the ML and the NJ tree. To better assess what observed by simply looking at the tree, we statistically compared models of character transitions. All biologically relevant models that we tested (the “oneway step-wise accumulation” model, the “free oneway acquisition” model, the “bidirectional step-wise” model, and the “free oneway loss” model; see Material and Methods for details) resulted less informative than both the ER, the SYM, and the ARD models. Among these, ARD resulted the best-fitting, therefore admitting a different rate for each possible transition among the G0, G1, and G2 states (predicted rates are summarized in Figure 7). Among the rates inferred by the model, the highest was represented by the G1→G0 transition, followed by G2→G0, therefore apparently contrasting the expectations. The rates of transitions from G0 to either G1 or G2 were lower than the specular transitions, but nevertheless the rate for G0→G1 was more than double the direct transition G0→G2, that was the transition with the lowest rate. Confirming previous tree topology considerations, the ASR performed with the ARD model of character transitions predicted G0 as the ancestral state in the most recent common ancestor of all Tudor domains with nearly 100% probability (Figure 7).

	🟨 G0	🟪 G1	🟩 G2
🟨 G0		0.0096	0.0039
🟪 G1	0.1027		0.0078
🟩 G2	0.0155	0.0050	

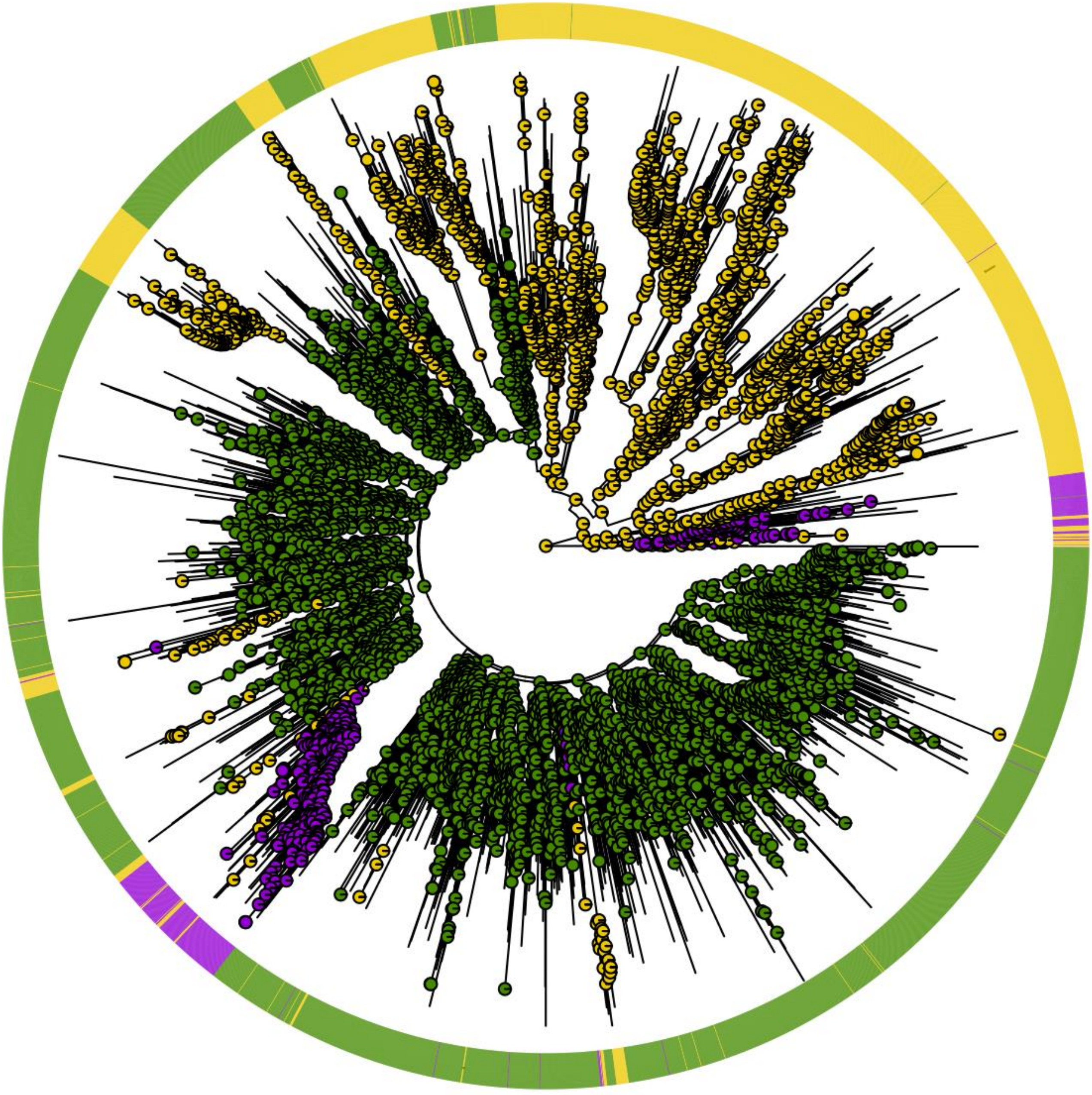


Figure 7. Ancestral State Reconstruction of structural N-t extensions. On top of figure is depicted the table of the transition rates as predicted by the *fitDiscrete* function on R following the ARD model (numbers refer to the transition from row to column). ASR was performed with *CorHMM* function on R following the ARD model. Nodes are coloured based on the predicted probability of each state (i.e. the three Tudor domain structural groups, G0-1-2; see transition rate table on top of figure for colours; yellow: G0; purple: G1; green: G2; colours as in Figure 3, 5, and 6). The states of the terminal tips of the tree are depicted in the outer circle that surrounds it. Outgroup is not present.

Discussion

Step-wise accumulation of N-t secondary structure could not be confidently confirmed, but the model remains a strong hypothesis

The fact that the Tudor domain secondary structure without N-t extensions represents the most ancestral condition was confirmed by our analysis. Indeed, all early branching clades (antecedent to Clade B of Figure 5) contained domains displaying exclusively this structure, and not surprisingly the ASR statistically confirmed that the stem node of Tudors was almost certainly G0. However, at present, the clear pathways of evolution of the secondary structure could not be confidently resolved because of the noisy ML tree. Consequently, the ASR and the inference of character transition rates (Figure 7), calculations that relied on the topology of the tree itself, cannot be considered completely reliable. Moreover, the tree presented here and used for the ASR inference was chosen among the 5 IQ-TREE replicates as the more biologically coherent, therefore these results can be considered circular. However, ASR analyses identified G0 as the most ancient state in 3 of the 5 trees and, in the remaining 2, G2 was considered as the basal state (with nearly 70% probability against 30% for G0) due to the presence of some scattered and disordered clades of basal G2 domains with no relationships to their belonging OG (most were mixed OG164 ones; data not shown). Moreover, the HMM profile tree undoubtedly showed how G2 domains were monophyletic and clustered together in a more derived position respect to all other G0 and G1 sequences (Figure 6).

Speculation on the subsequent evolution to a G1 domain with the N-t α-helix can be made from the observation that the OG including SMN1-2-DC1 clustered together with 10 basal sequences of TDRD3. Moreover, the earliest branching clade is represented by mixed SMN1-2-DC1, TDRD3 sequences, and in all other ML tree replicates domains belonging to these genes are consistently clustered together (data not shown). While all G0 domains of other proteins share methyl-lysine binding activities, TDRD3 is the only one that binds methylated arginines, like G1 and G2 domains (see Introduction). The observed phylogenetic relationship with the SMN-related Tudor domains suggests that these G1 domains might have evolved from the addition of an α-helix to a co-opted TDRD3 G0 domain. Another possibility could be that these proteins independently evolved methyl-arginine binding activities that led to the convergent fixation of similar amino acids in similar positions. The G1 domains of ALG13 clustered separately from them in the ML tree and this could mean that the common ancestor of the protein independently evolved the same N-t extension, instead of a common evolutionary origin. Unfortunately, lack of tree resolution and lack of precise binding activities of the ALG13 Tudor domain do not allow for a confident discrimination of the two scenarios, that remain both possible. Nonetheless, the HMM profile tree (Figure 6), where ALG13 and SMN-related domains cluster together, suggests a common origin for the same G1 N-t extension, even if the sister relationship with TDRD3 is lost (they are nevertheless close in the tree).

However, some contradictory results of the present ML phylogenetic tree are represented by the apparent reversals to a G0 or G1 state after the acquisition of the full set of N-t extensions (two β-strands and one α-helix, i.e. G2). This scenario cannot be excluded; however, some observations hold against it, or at least against the fact that it happened as frequently as the tree suggests. Indeed, all annotated G0 genes have been previously identified as shared eukaryotic genes, since they have all been found in non-metazoan lineages, while almost all G2 genes are metazoan innovations. Also in the present analysis, we could observe how non-metazoan Holozoa shared the presence of most G0 and G1 genes, while the only G2 gene present in their genomes was SND1. This different origin is reflected in the long stem branches of the monophyletic clades that include domains belonging to eukaryotic-wide genes. This is true for all G0 and G1 domains (including all those nested within the B clade) and also for the G2 domains of SND1, the only G2 gene that is not a metazoan-specific innovation. The fact that some eukaryotic-shared domains were clustered nested inside Clade B (that comprehended all G2 domain) is probably the result of a computational construct of tree inference. Indeed, it is less parsimonious to think that a domain organization that preceded the evolution of animals was substituted by newly evolved domain structures and lately re-evolved into the ancestral condition appearance. Indeed, the NJ of the alignment profiles showed the monophyly of all G2 domains in a derived position, with no G0 or G1 domains nested within, therefore further suggesting how the evolution of those N-t extensions was a single event and there were no back transitions of the secondary structure.

Also the proposed hypothesis of the G2 structure that evolved once from the insertion of a G1 domain within the SN domain of SND1 (Jin et al. 2009) could be neither confirmed nor excluded. SND1 domains formed a well-supported clade but their relationships with other domains were weak: no G1 nor G2 domains confidently clustered in sister-relationship with it. Also in the less noisy HMM profile tree (despite apparently confirming the evolution of the G2 structures as a single event), SND1 is nested between all other metazoan-specific Tudor domains. Given the lack of resolution from the present analysis, the fact that it was the only G2 gene that we could find in non-metazoan Holozoa support the hypothesis that the animal G2 radiation derived from co-option of its Tudor domain. However, the structural order of the secondary extensions from which the SND1-derived G2 evolution hypotheses stemmed, remains the strongest evidence (Jin et al. 2009).

In order to increase the confidence of the tree, and therefore of clade relationships and ASRs, different approaches can be conducted. Tree inference itself can be improved by performing many additional independent replicates of IQ-TREE, therefore exploring more pervasively the tree space and avoiding local likelihood peaks. Moreover, the domain set could be split in different subgroups (based *a priori* on OGs and taxonomic units, or *a posteriori* on the monophyletic groups retrieved by the whole tree) in order to obtain less noisy trees that would then be compared among each other to test the consistency of the predicted relationships. Alternatively, additional runs can be performed on representative subsamples of the sequences, therefore improving the ratio between the number of tips and the number of alignment positions, i.e. improving the resolution power and the inference confidence).

G0 and G1 group distributions are partially explained by differences in genomic architecture

We could observe significant and positive correlations between the G0 and G1 group distributions in the species, and the ratio between the number of genes and the number of proteins (Figure 4). Such ratio is relevant in terms of the molecular phenotypic complexity, since, while the number of genes between two genomes might be similar, the number of possible transcripts, and therefore of possible proteins, can widely vary. For instance, *Xenopus tropicalis* and *Dendronephthya gigantea* share an almost identical number of genes (21,898 and 22,045, respectively) but a 1.5-fold difference in the number of transcripts-proteins (45,171 and 28,741, respectively; numbers obtained from the present study; see Materials and Methods and Supplementary Table B2). A higher level of transcriptional potential intuitively should require a finer tuning of both transcriptional and post-transcriptional regulation. Indeed, the number of transcription factors in a genome increases more rapidly than the total number of genes (Nimwegen 2003). This suggests that with different genome complexities, the number of regulatory states change, and the more there are, the more transcription factors expand in a super-linear trend (Nimwegen 2003).

The correlations that we found with G0 and G1 Tudor groups might be the reflection of it. Tudor domains that belong to the G0 group are included in proteins that mostly have histone-binding activities and are involved in chromatin organization and gene regulation (see Introduction). On the other hand, SMN1, SMN2, and SMNDC1*,* that are included in one of the two annotated OGs that fall in the G1 group, are known to be core components of the SMN complex involved in spliceosomal small nuclear ribonucleoprotein assembly (Kolb et al. 2007). The correlation of G0 and G1 groups with the ratio between the number of genes and the number of proteins is coherent with their functional annotations. Moreover, the G0 distribution is correlated with the gene density of the genome. Given the shared epigenetic regulative function of the proteins included in such group, a more numerous or more diverse set of G0 Tudor genes might be selected in species with “diluted” genomes, while on the other hand they could be redundant and eventually lost in compact genomes with close gene spacing.

However, the G0 Tudor group comprises a wide range of proteins that have diverse functions and it is not easy, and not properly correct, to consider them together as shaped by the same selective pressures. Moreover, many of these genes encode for proteins with key cellular roles, and their evolutionary dynamics most certainly cannot be reduced to few genomic features. On the other hand, the correlations are intriguing and somehow coherent with their functional landscape, and the fact that each of these proteins have been observed as completely lacking in at least one phylum (see Figure 1; e.g. the DNA-repair-associated TP53BP1 in Ctenophora, Rhombozoa, and Orthonectida – also lacking in *D. melanogaster*) suggests that organisms can nevertheless manage to function without them according to lineage-specific evolutions. Indeed, while for many of these proteins it was previously assessed the presence in many eukaryotic clades, some of them were completely lacking in all the unicellular holozoan phyla considered in the present analysis (see Figure 1; e.g. SETDB1 and UHRF1).

The evolutionary patterns of G2 Tudor genes are consistent with multicellularity-related molecular innovations in animals

As regards the G2 group a more direct association with function can be made. Indeed, protein belonging to this group are almost exclusively associated with the piRNA pathway of TE silencing in the germline (Jin et al. 2010; Ying and Chen 2012). Such pathway is fundamental for proper germline formation and multipotency maintenance, as confirmed by the strong conservation across animals of the proteins involved (Fierro-Constaìn et al. 2017). We indeed found a stronger correlation between the number of *piwi* homologues and that of G2 Tudor genes, in respect to the other two structural Tudor domain groups (Figure 4). The genes that ended-up in what we called the *piwi*-like OG were homologues of *piwi*, *ago3*, and *aub*, genes directly involved with key-roles in the germline-specific piRNA biogenesis (Czech et al. 2018). However, we could not find any correlation with *ago*-related genes and with other Piwi domain-containing genes, confirming how the strict relationship with *piwi*-like proteins is specific and limited to these, rather than to generic Piwi-containing proteins.

Piwi, together with other GMP genes (like *vasa* and *nanos*; see Chapter A), are thought to be metazoan innovations (Ailé et al. 2015; Fierro-Constaìn et al. 2017) and they probably evolved alongside multicellularity. Indeed, the evolution of the germline was sided by the evolution of numerous innovative genetic features, such as the aforementioned GMP genes. Some of the Tudor genes, namely those belonging to the G2 structural group (except for SND1), are comprised among these innovations (Fierro-Constaìn et al. 2017) and in fact the correlation with the *piwi*-like gene distribution throughout animal species highlighted the intimate molecular relationship that ties them together. We could also confirm the animal-wide distribution of these genes, that could be found in the present analysis also for the basal clades of Porifera, Cnidaria, and Ctenophora. Moreover, also

STK31, that was considered as a bilaterian-specific innovation (Fierro-Constaìn et al. 2017), could be found in many Cnidaria and Ctenophora species, extending it to a metazoan-wide level.

Most G2 Tudor proteins were grouped within the large OG164 homology cluster, despite they belonged to different genes. These genes almost perfectly coincided with the Tudor genes that were considered as metazoan innovations involved in the GMP program by Fierro-Constaìn and colleagues (2017) (excluding TDRD9 and STK31 that constituted independent OGs). Curiously the relationships among them could not be confidently solved and they were ascribed in the same homology cluster. Such ambiguity was even more accentuated with the inference of the ML phylogenetic tree, since the Tudor domains of these proteins were scattered along the tree in an apparently chaotic manner. Our tree results must be taken with extreme caution because the inference of that many nodes is based on an alignment built on short sequences (domains). Nonetheless, for many of the considered Tudor domains, we managed to obtain monophyletic clades, whose relationships to one another were however difficult to confidently assess (Figure 5). For the domains belonging to OG164 this could not be neatly obtained, suggesting a convoluted pattern of evolution that involved multiple duplications and maybe also a bricolage-like evolution of the proteins. Indeed, while G0 and G1 domains are usually present in single copies within their protein, or at most in tandem couples, the G2 domains display an enormous range of different occurrences, from single copies to up to 19 multiple copies within the same protein (some OG164-included cnidarian sequences; 19 copies in *Orbicella faveolata*). In some cases, the copies resulted from within-protein duplications, hence they clustered together in the tree, but most of the times they did not. When this happened, one possibility is that exons coding for the domain shuffled among different genes. This is indeed a known way of protein architecture evolution that can be obtained in different ways, from recombination (homologous or nonhomologous) to action of mobile elements (see Forslund et al. 2019 for a review on evolution of domain architectures). Moreover, gene evolution by exon shuffling has also been associated to a higher frequency in Metazoa with respect to other eukaryotic lineages (Bjorklund et al. 2006; Ekman et al. 2007). The convoluted evolution of OG164 domains suggests that a complex and non-linear pattern of domain duplications and insertions within and among genes characterized the early evolution of the germline-specific Tudor proteins.

Another non-mutually exclusive explanation of the lack of tree resolution could lie in the relatively fast times of evolution of the set of OG164 germline-related proteins. Indeed, they are metazoan innovations, or at least metazoan radiations, related to the establishment of cellular lineages that arose early in multicellularity (that are germline and multipotent lineages). Their evolution might have been relatively fast, or at least fast enough to impede distinguishing their reciprocal evolutionary relationships after almost one billion years of sequence evolution (always considering the short length of the domain). Indeed, in our phylogenetic tree, while basal branches of clusters containing G0 Tudor domains were relatively long, G2 ones were short and lowly supported. These data support a fast lineage-specific ancient radiation of G2 that may not allow for a sufficient resolution. On the other hand, G0 are included in proteins shared by many eukaryotic lineages and their stem branches reflected the much longer evolutionary times.

Indeed, all G2 genes (with the exclusion of the anciently-evolved SND1) were completely lacking in unicellular Holozoa, together with sequences belonging to the *piwi*-like OG. There were only two notable exceptions to this pattern: the two Ichthyosporea *I. hoferi* and *Chromosphaera perkinsii*. The latter had an additional G2 domain in a sequence that clustered in a small OG together with a couple of Tardigrada; the former had a single G2 Tudor sequence clustered in the germline-related OG164. Curiously, these two species were the only ones that included also a gene in the *piwi*-like OG (with the precise domain architecture of *piwi*), despite homologues of such gene were never found outside Metazoa. These observations are undoubtedly intriguing, and it might suggest that at least the available genetic toolkit was shared by the common ancestor of all Holozoa. Then it might have been lost in most unicellular phyla while it expanded in Metazoa. However, most Ichthyosporea known so far are either parasites or symbionts of animals, and the presence of *piwi*-like and G2 Tudor genes in their genome could be also due to horizontal gene transfer from the host (however, *C. perkinsii* is, at least at its present evolutionary history, and at the state of our knowledge, a free-living organism). Be that as it may, sequence analyses on the sequences belonging to these two species are definitely worthwhile to discriminate between the two hypotheses, and results might provide important hints on the evolution of the germline-associated genetic toolkit.

Piwi loss and genomic dynamics as two driving forces of Tudor gene evolution: the cases of Tudor gene set reductions

Considering Metazoa, an immediate pattern that emerges from the lower edge of the distribution of Tudor genes in our species data set is the fact that 15 out of 17 species with less than 15 Tudor genes were endoparasites (Figures 2 and 3). Parasitism, and especially endoparasitism, usually leads to the evolution of reduced morphological complexity, that has also been associated to a reduction in genome size and/or gene composition (Poulin and Randhava 2013; Jackson 2014; Zarowiecki and Berriman 2014; Chang et al. 2015). Most of these conclusions have been, however, drawn for unicellular organisms (such as for the extremely reduced genomes of Microsporidia: Nakjang et al. 2013), and when analysing the independent evolution of parasitism in animals the situation was not neat. Data from independent occurrences of parasitism evolution in Nematoda, for instance, reveal how both the genome size and the gene content is highly variable and not straightforwardly associated to lifestyle strategies (Blaxter and Koutsovoulos 2014; Viney 2018).

The endoparasites of our data set are included in Nematoda, Platyhelminthes, Rhombozoa, Orthonectida, and Cnidaria, and represent at least 9 independent evolutions of endoparasitism (data from the survey of Weinstein and Kuris 2016): a single origin considered for the 3 species of Myxozoa, highly derived parasitic Cnidaria (*Thelohanellus kitauei*, *Myxobolus squamalis*, and *Henneguya salminicola*); an origin for the Orthonectida species *Intoshia linei*; an origin for the Rhombozoa *Dicyema japonicum*; a single origin for the 6 species of Platyhelminthes Neodermata (*Echinococcus granulosus*, *Schistosoma mansoni*, *Dibothriocephalus latus*, *Opisthorchis viverrini*, *Fasciola hepatica*, and *Protopolystoma xenopodis*); and 5 independent origins for the 6 species of Nematoda (*Trichinella spiralis*, *Necator americanus*, *Strongyloides ratti*, *Bursaphelenchus okinawaensis*, *B. malayi*, and *L. Loa*; the last two sharing a common origin of parasitism). The interesting pattern is that for 8 out of these 9 independent evolutions we could find a common pattern of strong reduction of Tudor genes (*B. malayi* and *L. loa* had a number of Tudor genes comparable to free-living species). However, it looks like this pattern is indeed strong but slightly different among the phyla, with different sets of Tudor genes that were lost, and we could hardly find a single interpretation coherent for all of them and exclusive for endoparasites.

Piwi-like gene loss explained the strong reduction of G2 Tudor genes in Neodermata. All of them lost *piwi* and its closely related homologues, as previously observed in other works (Tsai et al. 2013; Fontenla et al. 2021), bringing to the complete loss of the piRNA pathway and most of its associated genes (Fontenla et al. 2021). Indeed, in our survey, we could find only 2 G2 Tudor genes for each species: SND1 (not strictly involved in the piRNA pathway) and a single-Tudor OG with no homologues outside Platyhelminthes. Previous authors coupled the loss of *piwi* in these species to the expansion of a Neodermata-specific set of the Argonaute gene family (called FLAgos), proposing them as a novel adaptation to supply the loss of the canonical piRNA pathway (Skinner et al. 2014; Fontenla et al. 2021). We could assess that also *piwi*-associated Tudor genes have been lost and they were not co-opted in the putative novel molecular strategy, stimulating its further characterization.

Also the nematodes *B. malayi* and *L. loa* did not have *piwi* copies, but they were the only endoparasites that maintained an amount of Tudor genes comparable to other metazoans. When looking in detail their G2 Tudor gene composition, however, we could notice that they indeed lost all classic piRNA-associated G2 genes, and the G2 genes present were sequences containing exclusively the Tudor domain that were included in nematode-specific OGs or in non-annotated OGs shared with few other phyla. Interestingly, we could observe the same loss of canonical G2 genes in other Nematoda who also lost *piwi* (endoparasites *T. spiralis* and *S. ratti*, and the free-living *Aphelenchus avenae*), but also in species that did not lose it, like *B. okinawaensis*, *N. americanus* and the free-living *C. elegans*. Therefore, all nematodes of our dataset showed a reduction of the canonical germline-related Tudor gene set, despite the presence of *piwi* homologues in their genome.

The only exception was represented by *P. sambesii*, that underwent an expansion in both G0 and G2 Tudor genes, comprehending those belonging to the germline-associated OG164 (this species also presented two copies of *piwi* homologues). Beltran and colleagues (2019) identified two main mechanisms of piRNA organization and biogenesis in nematodes, and while *C. elegans* had the so-called C-type, *P. sambesii* had the P-type. It is intriguing to associate these two different strategies to the different signals that we observed in our analysis: both these species had *piwi* homologues, but while *C. elegans* lost most canonical G2 Tudor genes, for *P. sambesii* we could annotate 17 sequences in the germline-associated OG164. It might be possible that the two different piRNA organization and biogenesis mechanisms were associated to different *piwi*-related gene evolutions. Unfortunately, our data set did not contain other species analysed by Beltran and colleagues (2019), and more pervasive genetic investigations would be needed to strengthen this hypothesis, that so far is just a suggestion.

Loss of *piwi* was indeed previously observed and interpreted as a phenomenon that occurred independently in almost all non-clade V Nematoda lineages (but here also Clade V *S. ratti* was observed lacking *piwi*, and Clade IV *B. okinawaensis* was observed having one piwi homologue), and both nematode-specific and ancient non-metazoan eukaryotic strategies of TE silencing (involving chromatin remodelling and DNA methylation) were proposed as evolved alternatives (Sarkies et al. 2015). However, this suggests that Tudor gene loss is a pattern shared by Nematoda and not a parasitism-specific or a strictly *piwi*-related feature (see *C. elegans* and *B. okinawaensis*). Also, in previous works, different TE loads in Nematoda were not related to either life strategy or *piwi* loss, but mostly interpreted as the product of genetic drift (Szitenberg et al. 2016). Nevertheless, even if not free from exceptions, we could confirm an intimate relationship of G2 Tudors evolution to the piRNA pathway modification also in the Nematoda clade.

The loss of *piwi* can indeed explain some of the observed Tudor gene reductions in our dataset, but limited to G2 genes (not G0 and/or G1). Moreover, while usually *piwi* loss was sided by the loss of canonical G2 Tudors, the opposite was not always true, like the aforementioned nematode cases, or like in all Myxozoa and *I. linei* where *piwi* is present even if in these lineages the almost complete loss of G2 Tudor genes was observed.

As said before, parasitism has been often associated in a causal manner to a reduction of phenotypic/genotypic developmental and morphological complexity (Tsai et al. 2013; Jackson 2014; Zarowiecki and Berriman 2014). The genomic correlations that we could observe between the assembly length, the total number of genes, and the number of Tudor genes were partially coherent with this pattern. The highest levels of phenotypic complexity reduction in our data set were represented by the Orthonectida *I. linei*, by the Rhombozoa *D. japonicum,* and especially by Myxozoa, whose tremendous adaptation to parasitism led to the loss of most tissue and cell specifications, leaving a handful of cell types and unicellular life stages (Feist et al. 2015). Indeed, all these three phyla have been observed as lacking key genes and pathways related to metazoan development (Chang et al. 2015; Mikhailov et al. 2016; Zverkov et al. 2019): mere global gene content reduction was observed in our data set for all Myxozoa, *D. japonicum*, and *I. linei*. Tudor genes related to gene expression and chromatin regulation (G0), and those related with the piRNA pathway (G2) apparently followed the same destiny.

Following this trend, also for Neodermata flatworms we observed an overall reduction of gene content (compared to most other species of the dataset and to the free-living flatworms *Schmidtea mediterranea* and *M. lignano*; however, see below for a recent Whole Genome Duplication for the latter; pattern confirmed by Hahn et al. 2014). However, this was not true for all the species: *D. latus* had a number of genes comparable to other Metazoa, and *P. xenopodis* had almost twice as much genes as *S. mediterranea*. Nevertheless, both these species had underwent a massive reduction of Tudor gene set, that we therefore interpreted as mostly driven by the loss of the *piwi* pathway rather than genomic trends, as said before.

For Nematoda we could also observe a lower total number of genes for parasitic species compared to the free-living species *C. elegans, Aphelenchus avenae,* and *Plectus sambesii.* However, in these cases, the variability of the Tudor gene sets probably followed lineage-specific evolutionary pathways, as said before. Indeed, *C. elegans* itself is present in the lower part of the Tudor gene distribution of our study, and while 5 independently evolved parasitic nematodes experienced a reduction in respect to it, in Spirurida we could observe a Tudor gene expansion that was not related to gene content (*B. malayi* has the lowest number of total genes among nematodes of the present study). Nematoda genomes experienced high rates of gene loss and gene acquisition bringing to a high proportion of sequences with no homologues outside the phylum (Rodelsperger et al. 2013; Rodelsperger 2017). Such genomic evolution of the phylum might be sufficient to explain the patterns observed in our species data set (see for example the lineage-specific G2 genes discussed above), regardless of the parasitic habit. Coherently, different lineages lost different Tudor genes, making it difficult to advance generalizations on parasitism-related modifications. To summarize, the phenotypic, genomic, and lifestyle evolutionary history of Nematoda appear extremely convoluted, and no single consideration can be made to explain their extreme variability.

However, following the suggestion of phenotypic/genotypic/genomic complexity reduction, it is interesting to notice that the only two free-living species with less than 15 Tudor genes were the placozoan *Trichoplax adhaerens* and the urochordate *Oikopleura dioica*, that also have a high degree of body plan simplification. The former has very low levels of tissue differentiations and a simple life cycle (even if cryptic cellular complexity has been suggested: Srivastava et al. 2008), and the latter have a simplified Chordata body plan with a compact genome that lost entire gene networks involved in developmental regulation and epigenetic machinery (Ferràndez-Roldàn et al. 2019). Therefore, independently evolved phenotypic reduction not related to parasitism led to partially shared patterns of Tudor gene distributions, suggesting that similar data for endoparasites might be the indirect result of traits that are associated, but not limited, to such life strategy.

However, when looking in detail, the trend is not completely free from outliers. For instance, the myxozoan *T. kitauei* had a relatively low total number of genes (approximately 14,000), even if still comparable with other free-living animals. This number was ~2 times higher than other Myxozoa, but nevertheless *T. kitauei* was the one with less Tudor genes among them. Also, among Platyhelminthes, *D. latus* had a number of genes comparable to most Metazoa (approximately 20,000), but still it had a relatively lower number of Tudor genes, similar to other Neodermata, while *P. xenopodis* had twice as much genes as the free-living *S. mediterranea* but had the lowest number of Tudor genes of the whole species distribution (together with *D. japonicum*). Additionally, a low gene content was also observed in free-living arthropods such as *Apis mellifera* and *D. pteronyssinus,* therefore not exclusively in species that lost many Tudor genes. Indeed, the metazoan variability is so high that whit this sample size we can only limit to observe general patterns and trends, confident that no overall generalization can be representative of the whole species distribution.

Piwi loss and genomic dynamics as two driving forces of Tudor gene evolution: the cases of Tudor gene set expansions

The genomic-related driving force of the reduction of Tudor genes is apparently mirrored and confirmed in our data set also by the expansions observed for the upper edge of the distribution. The 4 species with the highest number of Tudor genes were the bdelloid rotifer *A. ricciae* (102 genes), the free-living flatworm *M. lignano* (97 genes), the sturgeon *A. ruthenus* (72 genes), and the free-living nematode *P. sambesii* (72 genes). One of the intuitive opposites of genome reduction is the expansion due to Whole Genome Duplication (WGD) events, and all these species underwent such kind of major evolutionary events. The only exception is represented by *P. sambesii*, for which no information about WGDs is available in the literature. However, the genome survey of the conspecific *Plectus murrayi* revealed a much lower gene number with respect to *S. sambesii* (~14.000 against ~40.000; Xue et al. 2021). This might be the evidence of a putative WGD with massive retention of gene copies that occurred in the latter. Following this suggestion, the number of *S. sambesii* genes is not approximately the double of *S. murrayi* because of the genome decay process that occurred in the latter due to its extremophile lifestyle (Xue et al. 2021), but it is worthwhile to notice that the number is approximately the double of *C. elegans*. On the other hand, as regards the other species with a high number of genes, evidences of WGDs are much more solid.

Acipenseriformes separated from Teleostei around 350 Mya (Hughes et al. 2018) and did not experience the teleost-specific WGD. However, a different WGD apparently occurred in this lineage and the assembly of *A. ruthenus* genomes revealed the conservation to present times of a high degree of both structural and functional tetraploidy (Cheng et al. 2019; Du et al. 2020). Such evolutionary transition largely explains the Tudor gene set of this species, since all the three Tudor groups (G0-1-2) are expanded to almost exactly twice the amount of other vertebrates included in the data set. Coherently, also the genomic total number of genes is higher than other Vertebrata.

This explanation is consistent also for *A. ricciae* and other Bdelloidea, since degenerate tetraploidy was already present before the divergence of bdelloid families and after divergence from other Rotifera (Hur et al. 2008). Indeed, also *Rotaria socialis* and *Didymodactylos carnosus* (the other two Bdelloidea) have a higher number of Tudor genes in respect to other species (7th and 9th positions in the distribution of our 93 species, respectively), and much higher than *Brachionus manjavacas* (Rotifera, Monogononta). The fact that *A. ricciae* has almost double the genes of other Bdelloidea could be due to artefactual issues in genome annotation: a recent genomic comparison between desiccating and non-desiccating rotifers that included both *A. ricciae* and *R. socialis* highlighted similar apparent asymmetries in gene content but the authors interpreted this result as due to very low levels of divergence among homologues in *R. socialis*, suggesting that they would often collapse in bioinformatic assembling (Nowell et al. 2018). However, they also found that 81% of *A. ricciae* genome sites were presumably in double copy (2-fold covered) in respect to the congeneric *A. vaga* but excluded additional WGD due to equal chromosome numbers (it could be nevertheless due to partial genome duplications or endopolyploidy; Nowell et al. 2018). A combination of these two observations probably led to *A. ricciae* expansion of Tudor genes in respect to other Bdelloidea in our dataset: either duplication patterns for *A. ricciae*, or masked gene copies in non-*Adineta* bdelloids.

Also *M. lignano* (Platyhelminthes) recently underwent a WGD followed by the fusion of a whole duplicated set of chromosomes of the ancestral karyotype into a single large additional chromosome, leading to hidden tetraploidy (Zadesenets et al. 2017a-b). This partially explains our observed data on Tudor gene numbers, but not entirely. In fact, while G0 and G1 Tudor genes are nevertheless within the average of other species, G2 ones are significantly higher, being the most numerous G2 of all our data set species (63). Even considering the retention of the whole duplicated set, therefore dividing by 2 for simulating pre-WGD condition, the number is still high and would be among the first 10 animals of the distribution, calling the need for further interpretations.

G2 Tudor genes, as said before, are involved and evolutionarily tied to the piRNA pathway. This strategy of retrotransposon silencing is fundamental for animals especially in the germline, leading to infertility and absence of proper germ cell differentiation in the absence of molecular factors involved in it (Juliano et al. 2010; Siomi et al. 2011). However, it is essential for multipotency in

general, comprising also stem cell lineages (see Introduction and Chapter A). Free-living flatworms, such as *M. lignano,* display one of the highest levels of tissue regeneration potential in the animal kingdom thanks to totipotent cells called neoblasts, that have been observed to express GMP genes like *vasa* and *piwi* (see Chapter A). These cells are fundamental for regeneration, tissue homeostasis, in some species also for asexual reproduction, and are extremely numerous in adult tissues, comprising up to 30% of the whole cell population of the adult (Sasidharan et al. 2013). The piRNA pathway of retrotransposon silencing in these species is therefore extremely important for the survival of the organism (*piwi* homologues mutants of *S. mediterranea* show similar phenotypes to lethally irradiated samples where neoblasts are completely depleted; Kim et al. 2020). In our analysis we could annotate only 3 *piwi* homologues for *S. mediterranea* (more than the median in Metazoa), but 14 sequences for *M. lignano*. Our observed increase in *M. lignano* in the number of G2 genes (19 belonging to OG164, 4 to TDRD12, 3 to TDRD9, and 36 belonging to non-annotated OGs) could be due to the expansion of the piRNA pathway that happened in this clade due to the key and constant role of it in adult survival. Indeed, Long Retrotransposable Regions (LTRs), whose mobility is directly controlled by piRNAs, consist of 21% of *M. lignano* genome (Wudarski et al. 2017), and also in *S. mediterranea* 29% of the genome is composed of retrotransposon, including three families of enormous, possibly active, >30kb LTRs (Grohme et al. 2018). Data on their activity are lacking, but the observed expansion of *piwi* and G2 Tudor genes in *M. lignano* might be indirect evidence of it.
The activity of TE elements might have selected for the maintenance of G2 Tudor genes also in Bdelloidea. Indeed, also in these 3 species, the proportion of G2 genes in respect to other Tudor groups is higher than the majority of other animal taxa. Curiously, an even higher G2-biased proportion is present in the springtail *Folsomia candida*, the 6th species for number of Tudor genes of our data set (59 Tudor genes, of which 45 of G2 group). Bdelloidea and *F. candida* are all apomictic parthenogenetic species, and it might be tempting to associate the G2-Tudor expansion to the selection for an efficient TE control pathway in species that lack sex-related defenses to it (like genetic exchange among individuals and meiotic recombination). The consequences of asexuality and unisexuality on TE loads have always been of interest. It has been predicted that asexual populations that lacked efficient ways for controlling TE expansions should either accumulate them up to lineage extinction or lack TEs in the first place (Dolgin and Charlesworth 2006), suggesting that TE content should be low in these species. However, this prediction was not always confirmed, and asexual or unisexual lineages are not usually characterized by different TE loads with respect to closely related sexual ones (see for instance Kraaijeveld et al. 2012; Bast et al. 2016; but see Jaron et al. 2021). In long-term asexual Bdelloidea, for example, relatively abundant, diversified, and recently active transposons and retrotransposons have been found (Nowell et al. 2021; but see also the recent LTR expansion observed in *A. vaga*: Kim et al. 2018). The same authors did not find any significant

difference in respect to other Rotifera in terms of TE load, but they however found bdelloid-specific expansions of TE silencing pathways (Nowell et al. 2021). In our species we could indeed confirm a higher proportion of G2 Tudor genes in respect to other Tudor groups, and a relatively high number of *piwi* homologues (8 to 9, against an overall metazoan median of 3), suggesting that this could have been selected to avoid detrimental effects due to TE mobility in their genomes in the absence of other effective molecular mechanisms of TE dynamic prevention. Also the parthenogenetic springtail *F. candida*, coherently, in addition to display a high number of G2 Tudor genes, also has the highest number of *piwi* homologues of our data set (19 genes), but data on TE activity are lacking.

Conclusions

In the present analysis we investigated the evolutionary pathway of the Tudor domain in Holozoa, and its distribution in the genes of 111 species. We could assess the widely diffused presence of Tudor genes in all phyla, and a notable expansion of G2-type ones in animals, confirming that the early evolution of Metazoa was sided by a relatively fast expansion of such gene family. This was driven by the convoluted and bricolage-like evolution of different Tudor genes involved in the germline/multipotency molecular pathway of retrotransposon silencing through piRNAs.

By looking at the distribution of Tudor genes in extant animal species, we could assess how the evolutionary dynamics of *piwi*-like genes (the key factors of the piRNA pathway) can largely explain the patterns of germline-related Tudor gene distributions, confirming their intimate and almost exclusive molecular relationship. However, also more general genomic evolutionary trends, such as genome simplification and genome duplications, can explain reductions and expansions of the Tudor gene set. Despite tempted by the observation that most of the species with low numbers of Tudor genes were endoparasites, we invoke caution in interpreting it as a direct causal connection. Some shared genomic patterns are indeed shared by these animals, and the reductions of the Tudor gene set observed in some taxa might be interpreted in the light of this trend. However, these dynamics are shared by these species, but not limited to them. Such a diverse and complex investigational unit (composed by specific high-order taxon traits and genomic/genetic adaptations) needs to combine different perspectives and interpretations, and general considerations might be led to stand on wobbling floors. Additional analyses with a higher number of parasitic organisms and including TE characterization and activity might contribute to provide additional suggestions.

Note:

The results exposed in the present chapter are currently being elaborated and integrated in sight of the submission to a journal with IF in the near future.

Chapter C

Germline Differentiation in Bivalves: TDRD7 as a Candidate Factor Involved in *Ruditapes philippinarum* Germ Granule Assembly

Introduction

Germ cells play a unique role in animal heredity and evolution as carriers of the genetic information across generations. In Metazoa, the hereditary information moves in two ways: within the germline (that in sexual animals eventually produces gametes), providing an immortal link to the next generation, and from germ cells to somatic cells to build a new organism. Therefore, investigating timing and mechanisms involved is a central challenge for understanding the evolutionary origin and maintenance of the crucial differentiation of the two lineages.

Solana (2013) introduced the concept of PriSCs (see Chapter A). These cells are evolutionarily conserved stem cells that act as a link of germline determinant expression from the zygote to the future germline (Solana 2013). PriSCs share common features with stem cells thanks to their capacity to self-renew and differentiate into specialized cells that can have both somatic and germline potential (Xie and Spradling 2000; Li and Xie 2005; Solana 2013). At some point during development, or several times during the life of animals with gonad renewal or high regeneration potential, a PriSC gives rise to a new PriSC and a Primordial Germ Cell (PGC) through an asymmetric cell division. PGCs are cells in proliferative state that retain self-renewal capacities and give rise to cells with only germline potential. PGCs will then populate the gonads through mitotic proliferation and give rise to germ cells and gametes by meiosis.

In the past decades, various research teams have focused on identifying and characterizing the cells that act as a link between zygote and gametes. Extensive research has focused on the identification in model organisms of germline determinant transcripts/proteins, many of which appear to be evolutionarily conserved through Metazoa, both for their presence in the genome and for their germline-related functions (the GMP, e.g. *vasa*, *nanos*, *piwi*, and Tudor genes; see Chapter A and B; Juliano et al. 2010; Ewen-Campen et al. 2010; Fierro-Constaìn et al. 2017). The timing of their expression and the level of organization at which they cluster together forming germ plasm, or germ plasm-related structures, is extremely variable in different animals (Kloc et al. 2004; Whittle and

Extavour 2017). However, a typical feature of germ cells is the presence, at least in some point during their differentiation, of germline determinants assembled in a differentiated region of the cytoplasm, generically called “germ plasm” (Kloc et al. 2004; Extavour 2007; Voronina et al. 2011; Solana 2013). These structures arise in specific stages of germ cell differentiation in some species, while in other they are present continuously throughout the whole germline, them being selectively inherited from the oocyte/zygote to a specific cell lineage. When the latter pattern is present, it is usually referred to as “preformation” (Extavour and Akam 2003). This is in contrast with germline specification by “epigenesis”, that involves the presence of germline-inductive signals from neighbouring cells surrounding the future germline during or after embryogenesis and it is thought to represent the ancestral mode of specification in Metazoa (Extavour and Akam 2003). However, beside the timing of their appearance and the level of involvement in germline specification, it appears that ribonucleoproteic cytoplasmic germ granules are present and fundamental for the functioning of germ cells in animals in general (Voronina et al. 2011; Sengupta and Boag 2012). Moreover, this is true also considering the general germline definition proposed by Solana (2013), i.e. including also totipotent cells formerly considered somatic: for instance, planarian neoblasts have perinuclear chromatoid bodies containing GMP elements that disappear in the differentiated progeny (Krishna et al. 2019). For this reason, germ plasm/granules characterization and study are crucial for the understanding of metazoan germline patterning.

Important factors acting in germ plasm assembly of model organisms are: Oskar (in polar granules of holometabolous insects; Ephrussi et al. 1991), Xvelo (in the Balbiani body of *Xenopus laevis*; Boke et al. 2016) and Bucky ball (orthologue of Xvelo in zebrafish; Bontems et al. 2009) being the most studied so far. Indeed, the function of the short isoform of Oskar (there are two isoforms that differ for 139 amino acids on the N-t) is to promote the formation of germ plasm, being necessary and sufficient for its assembly (Jeske et al. 2015), and it acts in concert with other components such as Vasa, Nanos, Piwi, and Tudor (Anne 2010). The Oskar protein is assessed to be present only in the insect lineage (Ewen-Campen et al. 2012), and it includes two functional domains: the RNA-binding domain Oskar, and the Lotus domain. While the former has been found so far only in Oskar proteins of insects and in Bacteria (the presence in insects is likely the result of horizontal gene transfer: Blondel et al. 2020), the latter can be found also in other germline-related proteins, such as homologues of TDRD5 and TDRD7 (Anantharaman et al. 2010; Callebaut et al. 2010). In some recent studies, it has been demonstrated that the Lotus domain of Oskar is responsible for the dimerization of the protein, and it physically interacts with Vasa to regulate Vasa DEAD-box helicase activity and to mediate its localisation in the germ plasm (Anne 2010; Jeske et al. 2017). Indeed, other indirect evidence of the Lotus-Vasa interaction come from other Lotus-containing proteins, such as the mouse TDRD7 and the homologues of TDRD5 and TDRD7 of *Drosophila* (Tejas and Tapas, respectively), that have been observed to co-precipitate with Vasa along with other germline components (Hosokawa et al. 2007; Patil et al. 2014). Indeed, these two Lotus-Tudor-containing proteins have been associated to roles in the proper assembly of cytoplasmic structures/granules in different species and in different tissues: from somatic ribonucleoproteic granules involved in the formation of ocular lens in mammals (TDRD7; Lachke et al. 2011), to chromatoid bodies in mammal male germ cells (TDRD7 and TDRD5; Tanaka et al. 2011; Yabuta et al. 2011), *Drosophila* germline perinuclear nuage (Tejas and Tapas; Patil et al. 2014), and granular structures in *D. rerio* germ cells (TDRD7; Strasser et al. 2008; D'Orazio et al. 2020). For these reasons, the presence of the Lotus domain in a protein might be a good starting point to characterize its functions within the germline and to try to predict Oskar-like germ plasm or germ granule assembly factors in other species that lack an identified master regulator, i.e. a factor that is necessary and sufficient for the assembly.

In our study, we approached the question in the bivalve *Ruditapes philippinarum*, an interesting developmental model. Beside an unusual modality of cytoplasmic inheritance known as Doubly Uniparental Inheritance (DUI) of mitochondria (Zouros et al. 1994; Milani et al. 2011) that makes it a unique and evolutionary stable study system for mitochondrial biology and inheritance, heteroplasmy, mito-nuclear coevolution and genomic conflicts (Breton et al. 2007; Milani and Ghiselli 2015; Ladoukakis and Zouros 2017), *R. philippinarum* shares with many other bivalves the annual renewal of gonads (Gosling 2003). Indeed, *R. philippinarum* gonads form every year at the beginning of the mating season. The gametogenic phase consists in the multi-step differentiation of germ cells inside sack-like structures, called acini, and leads to the ripening of the gonad. During this phase, the gonadic tissue develops inside the connective tissue, near the intestine, and consists of acini that grow in dimension with the progress of gametogenesis (Devauchelle 1990; Gosling 2003; Milani et al. 2011). After the spawning period, clams are characterized by sexual rest, gonads are degraded, and sexes are no more recognizable.

The annual gonadic renewal appears to be preceded by proliferation in the intestinal epithelium of undifferentiated cells that express germline markers, like Vasph, the *R. philippinarum* Vasa orthologue (Milani et al. 2015, 2018). Similar Vasa-tagged intestinal cell clusters were observed also in other bivalve species, such as the heterodont *Mya arenaria*, suggesting that it might be a shared pattern (Milani et al. 2017). However, similar reports lack from other species: studies in *Crassostrea gigas, Mytilus galloprovincialis*, and *Mizuhopecten yessoensis* do not discuss labelling of the Vasa homologue in similar cells, but only refer to Vasa antibody reaction in germ cells in the gonads (Obata et al. 2010; Cherif-Feildel et al. 2019; Mokrina et al. 2021).

It is clear how the characterization of the early germline stages in bivalves needs additional investigation, and the extensive diversity of the class represents a stimulating resource. Nevertheless, the annual renewal of the gonads is a shared characteristic and it would be interesting to understand how determinants are initially segregated into the germline and how germline continuity is preserved by specific cells during the non-reproductive season. Despite germline specification mechanisms in clams are far from being understood, recent analyses showed the presence of germ plasm related granules in *R. philippinarum* germline (Reunov et al. 2019). In that work, such granules, that include Vasph-positive substance, have been observed through Transmission Electron Microscopy (TEM) in early differentiating germ cells, i.e. spermatogonia and oogonia, associating them to meiosis onset. Then, during oogenesis, granular Vasph-tagged substance was observed to arise again at least twice: first, in the first stages of oocyte growth, and then in the late mature oocytes. These latter granules have been proposed to be selectively inherited in the germline lineage of the offspring, therefore determining it in a "preformation" mode of germline specification (Milani et al. 2018; Reunov et al. 2019). Indeed, early germ cells of both sexes show the presence of Vasph-tagged germ granules, that during the specification of the lineage dissolve and, in concert with mitochondria, appear to induce the mitosis-meiosis transition of spermatogonia and oogonia (Reunov et al. 2019).

The aim of the present study is to provide a better characterization of the germline formation in *R. philippinarum*. In this work, we explored the dynamics of germline development by *in silico* identification and *in situ* localisations of a newly identified germline marker for *R. philippinarum*. In details, starting from bioinformatic analyses on RNA-Seq transcriptomic data, we found a candidate possibly involved in germline differentiation in *R. philippinarum* (TDRD7 orthologue). We confirmed the *in silico* assembled sequence by Sanger sequencing, and we designed specific antibodies to target the protein *in situ*. Immunohistochemistry and immunofluorescence assays were used to study the distribution of TDRD7 within histological samples containing gonadic tissue. These experiments were performed on male and female individuals collected during the reproductive season and during the sexual rest.

Materials and Methods

Sequence identification and analyses

To look for potential germ plasm regulators in the Manila clam *R. philippinarum*, we started by BLASTing (Camacho et al. 2009) the *D. melanogaster* Oskar protein sequence (short isoform, i.e. the one that promotes germ plasm assembly) against the publicly available annotated bivalve proteomes (on the NCBI nonredundant protein database, or nr; taxid: 6544). We then used the best and only hit (*Mizuhopecten yessoensis* TDRD7A-like protein, accession code: XP_021379223.1) to look for the orthologue in *R. philippinarum* by BLASTing it against our *de novo* transcriptome. The *R. philippinarum* best hit was then back-BLASTed against *M. yessoensis* proteome (on nr; taxid:

6573) to assess orthology (them being reciprocal best hits). The *R. philippinarum* transcriptome we used was built with Trinity v2.9.0 (Grabherr et al. 2013) on reads from gonads and somatic tissues (abductor muscle and mantle) of 8 female and 8 male samples, and consisted in 553,711 transcripts (N50: 1,337; high number of transcripts is likely due to samples polymorphisms). The transcriptomic samples were part of transcriptomic profile analyses of *R. philippinarum* (NCBI BioProject Acc. No. PRJNA672267; same project used for Chapter A): total RNA was extracted with TRIzol, poly-A transcripts were isolated with magnetic beads and used as templates for cDNA synthesis; the selected insert size was approximately 500 bp; and sequencing was performed on an Illumina HiSeq 2500 platform to generate 150 bp paired-end reads. Reads were previously trimmed with Trimmomatic v0.39 (Bolger et al. 2014) with the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:28:28 MINLEN:98, leaving approximately 171 million reads for Trinity v2.9.0 *de novo* assembly. Transcriptomic completeness was assessed with BUSCO through the gVolante online interface (percentage of complete core orthologues: 99.8%; https://gvolante.riken.jp/analysis.html).

Once we obtained the *R. philippinarum* orthologue transcript of the *M. yessoensis* TDRD7A-like protein, we extracted the translated coding sequence with NCBI ORFinder (https://www.ncbi.nlm.nih.gov/orffinder/). The sequence was then characterized for domain composition (with InterProScan v5.45.80; Jones et al. 2014). Over the nucleotide sequence of the whole transcript, we designed 13 couples of PCR primers that covered the whole coding sequence in 9 overlapping sections (predicted with Primer3; Untergasser et al. 2012; Supplementary Table C1). We used the primers to amplify the transcript portions and Sanger-sequence them to verify the existence of the whole transcript *in vivo*, therefore excluding the possibility of it being a *de novo* assembly construct. Samples for amplification came from two female gonad samples of *R. philippinarum* (RNA extraction by TRIzol from fresh tissues, Thermo Fisher Scientific; retrotranscription with SuperScriptIV, Thermo Fisher Scientific; PCR cycles in Supplementary Table C1).

We also calculated and compared the levels of transcription of the *tdrd7* transcript throughout the samples. Transcript quantification was performed with Salmon v1.3.0 (Patro et al. 2017) and differential expression analysis with DESeq2 (Love et al. 2014). We compared *tdrd7* transcript expression between different tissues (somatic and gonads) and different sexes. We also compared the results with those of the differential transcription of the germline marker *vasph* (NCBI accession code: JO110167.1).

Sampling

We performed histochemical analyses on specimens of *R. philippinarum* from Sacca di Goro (Adriatic Sea, Ferrara, Italy), sampled in the gametogenic stage (May to July), and in the reproductive spent phase (November and February). The sex of gametogenic clams was determined via gametic smear observation under an optical microscope. The direct observation of gametes also allowed us to access the reproductive stage of the samples, that can be only hypothesized before the sampling due to possible significant yearly environmental variations. Whole gonads and parts of the digestive tube were either dissected and directly processed for Immunofluorescence (IF) and Immunohistochemistry (IHC) or stored at -80 °C for Western Blot (WB) analysis. In clams sampled during the reproductive spent phase (November and February) the entire body was dissected due to their tiny size and the difficulties of determining their sex.

Primary antibodies

We decided to investigate TDRD7 in tissues in the form of protein sequence (rather than mRNA localisation) because we were interested in the stages of actual functional expression of the factor, and data would have been more directly comparable with previous works on Vasph protein detection in the same tissues (Milani et al. 2017; 2018). In order to visualize TDRD7 in *R. philippinarum* we used four different antibodies that differed in the target sequence of the protein. Two of these antibodies were *ad hoc* produced based on the *R. philippinarum* TDRD7 sequences, while two were commercially available antibodies produced on the *H. sapiens* sequence.

For the clam-specific antibodies we utilized specific antisera produced in chicken by Davids Biotechnologie (Regensburg, Germany). These antibodies were generated against two synthetic peptides: the first one was synthesized from the 1st of the two predicted Lotus domains present in the TDRD7 protein (peptide EKFILSMPDVARIDRRGGD, acronym EKF), while the other was synthesized from the 2nd of the three predicted Tudor domains (peptide AYDDGLYHRVRVMSVQDGKK, acronym AYD). The peptides were chosen among those with better score for epitope prediction (algorithm by Davids Biotechnologie). Moreover, we evaluated the position of the suggested peptides in the 3D structure predicted on the I-TASSER server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/; Yang and Zhang 2015) and we chose external and easily reachable targets. The obtained antibodies were tested for immunoreactivity by ELISA with the immunogen peptides and were later purified by affinity chromatography (Davids Biotechnologie). Davids Biotecnologie also provided the synthetic peptides which were used for the primary antibodies production and that we used to test antibody specificity in the Western Blot assays.

The second set of polyclonal antibodies were commercially produced in rabbit by Abcam against human TDRD7 (Cambridge, United Kingdom). Ab241349 (acronym Ab49) was tested for WB in human cell lysates and the immunogen was human TDRD7 amino acids 1048-1098, corresponding to the C-t end of the protein. Ab224462 (acronym Ab62) was tested for tissue essays and the immunogen was human TDRD7 amino acids 750-950, that represents the inter-domain region between the 2nd and 3rd Tudor domains, partially overlapping with the C-t of the 2nd (but not comprising the epitope of anti-AYD) and the N-t of the 3rd.

Immunolocalisation

Females and males of *R. philippinarum* were analysed at two stages of the reproductive cycle (gametogenic and spent phase) to identify the localisation of TDRD7 protein in several tissues and cell types. The histological districts observed included germline (acini in gametogenic individuals) and somatic tissues (intestinal epithelium and connective tissue). Samples were processed with IF and IHC protocols.

IF For the IF protocol (Milani et al. 2015), samples (20 samples: 9 gametogenic ones per sex, and 4 in the spent phase) were fixed in a solution consisting of 3.7% paraformaldehyde, 0.25% glutaraldehyde and PIPES 2X (pH 7.0-7.2) for about 3 h and 30 min at Room Temperature (RT). Samples were washed with PBS (pH 7.2) for 1 hour with changes every 10 or 15 minutes. Then, samples were embedded in 7% agar and processed with a vibratome (Leica VT1000 S) to obtain sections of 150 µm thickness. Afterwards, the sections were dehydrated with increasing concentrations of methanol (50, 75, 80, 90, 100% for 10 minutes each) and rehydrated in TBS (pH 7.2) for about 1 hour. The sections were treated with sodium borohydride in TBS (pH 7.4) for a 1 h and 30 min at RT, and, after TBS washing, antigenic sites were exposed during 18 minutes of 0.01% Pronase in PBS. After permeabilization, non-specific protein binding sites were blocked with 10% Normal Goat Serum (NGS) and 1% Bovine Serum Albumin (BSA) in TBS-0.1%Tween-20 (Sigma) (pH 7.2; TBS-Tw) for 1 h and 30 min at RT, and section were ready for antibody reactions.

They were incubated with the primary antibodies (anti-EKF, anti-AYD: diluted 1:1,000; Ab49, Ab62: diluted 1:100; in a solution of 3% BSA in TBS-Tw, pH 7.4) for ~60 h at 4 °C, to which followed incubation with the secondary antibody for ~25 h at 4 °C (anti-chicken Dylight®550 Cross-Adsorbed, Thermo Fisher: diluted 1:800; anti-rabbit AlexaFlour®488, Thermo Fisher: diluted 1:400; dilutions in a solution of 1% NGS and 1% BSA in TBS-Tw, pH 7.2). Negative controls for the specificity of immunostaining were obtained by omission of the primary antibodies, replaced by 1% normal goat serum and 3% bovine serum albumin. TO-PRO3 nuclear dye (1 mM, diluted 1:1000 in PBS), with excitation wavelength at 642 nm and emission at 661 nm, was used for nuclear staining (10 minute-incubation). Sections were mounted in anti-fade medium (2.5% 1,4-diazabicyclo[2.2.2]octane, DABCO, Sigma; 50 mM Tris; and 90% glycerol) on the slides and stored at 4°C in the dark. Images of IF staining were acquired by confocal laser scanning microscope (Leica confocal SP2 microscope;

Leica Microscope Objective HCX PL APO 63x/1.32-0.6 Oil CS; image dimension: 1140x968 pixels) using Leica software. Fluorophores used were DyLight®550 (Ex: 562 nm, Em: 576 nm), AlexaFlour®488 (Ex: 495 nm, Em: 519 nm), and TO-PRO3 for nucleic acids (Ex: 642 nm, Em: 661 nm).

IHC For IHC, the entire body was processed following the method used by Lazzari and colleagues (2014). The samples (12 samples: 5 gametogenic per sex, and 2 in the spent phase) were fixed in a modified Bouin's fixative solution composed of saturated aqueous solution of picric acid and formalin (ratio 3:1) for 24 h at RT. After prolonged washing in 0.1 M PBS at RT with changes every 15 min, the specimens were dehydrated in graded series of ethanol (70, 80, 95, and twice in 100%, 10 min each). To facilitate solvent-ethanol replacement, the specimens were placed 2 h in the clarifying solvent Noxyl at 37 °C and periodically shaken. After two passes in melted Paraplast plus (Sherwood Medical, St. Louis) each for 1 h at 60 °C, the specimens were included in the Paraplast plus. Sections 5 µm thick were cut with a microtome (Leica RM2145), then mounted on silane-coated slides (Sigma) and dried. Then, sections were pre-treated in a stove at 55 °C overnight or 2 h at 60 °C. In order to clean the samples, paraffin was washed out with Xilolo I and then Xilolo II for 20 min and then then sections were rehydrated in a graded series of ethanol (twice in 100, then 95, 80, 70, 50% and water, for 5 min each). Endogenous peroxidase activity was blocked with 1% hydrogen peroxide (H_2O_2) in 0.01 M PBS (pH 7.4) for 25 min at RT, avoiding nonspecific background colour, and then washed in PBS 0.01 M for 10 min.

For the antigen retrieval, tissue sections were immersed in 0.01 M citrate buffer, pH 6.0 and heated in a microwave oven (750 W) for two cycles of 5 min each. The slides were cooled, then washed with PBS. The outline of the sections was drawn with a PapPen. Preincubation was performed in PBS containing 10% NGS, 1% BSA and 0.1% Tween-20 (Tw) for 1 h. The sections were incubated over night at 4 °C with primary antibodies (polyclonal anti-EKF or anti-AYD developed in chicken; polyclonal Ab49 or Ab62 developed in rabbit; diluted 1:100 in a solution of 0.01 M PBS containing 2% NGS, 1% BSA and 0.1% Tw) in a moist chamber on a floating plate. After washing in PBS with 0.1% Tw, sections were incubated for 1 h with secondary antibodies (HRP anti-chicken in goat and HRP anti-rabbit in goat, Santa Cruz Biotechnology Inc.; diluted 1:100). After rinsing in PBS with 0.1% Tw many times, the immunoreaction was visualized with diaminobenzidine (DAB), that, reacting with the peroxidase conjugated with the secondary antibody, forms an insoluble precipitate visible under an optical microscope. Sections were dehydrated in ethanol, cleared in xylene, and coverslipped with Permount (Fisher Scientific, Pittsburgh, PA). Negative controls for the specificity of immunostaining were obtained by omission of the primary antibodies, replaced by 3% normal goat serum. IHC imaging was performed with Olympus BH-2 microscope (Olympus S Plan Achromatic objective 10x, numerical aperture 0.30, working distance 7.50 mm, focal length 18.98 mm, and

Olympus S Plan Achromatic objective 20x, numerical aperture 0.46, working distance 1.50 mm, focal length 8.03 mm; both with Tube length/coverslip thickness 160/0.17 mm). Images were acquired with BEL Photonics BlackL 5000 USB digital camera (5 Mpixel) through the acquisition software BEL Photonics Eurisko 2.9 (auto exposure, 8-bit RGB images recorded in TIFF format, 14.2 MB in size, pixel dimensions: 2592x1920).

Western blot

WBs were carried out following the method used by Milani and colleagues (2015). Gametogenic male and female clams samples in July were used to obtain gonadic homogenates (14 samples: 7 per sex) and female gonads were freshly dissected and homogenized using an Ultra Turrax T25 (Janke & Kunkel IKA-labortechnik) in a buffer containing 10 mM Tris-HCl (pH 7.5), 1 mM ethylene glycol-bis(2-aminoethylether)-N,N,N',N'-tetraacetic acid (EGTA), and 0.1% Sodium Dodecyl Sulfate (SDS). To limit degradation, one protease inhibitor cocktail tablet (Complete Mini, Roche) and 1 mM PMSF were added to 5 mL of the homogenization buffer. Then samples were centrifuged at 7,500 xg for 10 min at 4°C. Supernatant was stored at -80 °C. The amount of total proteins in the homogenates was quantified with Lowry method (Lowry et al. 1951). Then 20-40 µg of total protein homogenate per lane, mixed with Laemmli Sample Buffer, was separated via 8.5% SDS-PolyAcrylamide Gel Electrophoresis (SDS-PAGE). Some gel lanes were cut and stained with Coomassie Brilliant Blue for the visualization of all the protein bands present in the homogenate.

For immunoblotting, proteins were electrically transferred to nitrocellulose membranes and, to evaluate the actual transfer of proteins, membranes were stained with Ponceau. Unspecific sites were blocked with 5% dried skimmed milk (Bio-Rad Laboratories, Hercules, CA, USA) and 3% Bovine Serum Albumin (BSA) in TBS-Tw for 1 h 30 min at RT, and then with TBS-Tw. The membranes were incubated with polyclonal primary antibodies (anti-EKF, anti-AYD, Ab49, and Ab62; diluted 1:1,000). To monitor the antibody specificity of the two clam-specific antibodies, synthetic peptides were incubated for 30 min with the primary antibody solution at a 20-fold excess concentration before use. The membranes were incubated overnight at 4 °C, then for 1 h at RT and later rinsed with TBS-Tw for 30 minutes. After washes, membranes were incubated with secondary antibody (HRP anti-chicken for anti-EKF and anti-AYD; HRP anti-rabbit in goat for Ab49 and Ab62; diluted 1:5000 in TBS-Tw 0.1% for 1 h at RT. The membranes were washed again for 30 min and the reaction was detected in a dark room using ECL Western Blotting Detection Reagents (Roche) and exposed to Hyperfilm ECL. Photographic plates were impressed for about 1 min and the blots were developed dipping the plates in developer and fixer solutions (Kodak professional).

Results

TDRD7 in *Ruditapes philippinarum*

The Oskar BLASTP against Bivalvia subset on NCBI nr database gave as best output hit *M. yessoensis* TDRD7A-like protein (XP_021379223.1; bitscore: 48.9; e-value: 6e-05; identity percentage of alignment region: 30.23%). The two proteins aligned exclusively in the amino acidic positions 18-98 on the Oskar sequence, that coincided with the Lotus domain (Oskar only Lotus: position 14-83 as inferred by InterProScan; TDRD7A-like first Lotus out of two: positions 3-76 as inferred by InterProScan). Indeed, Oskar has no homologues outside the Insecta lineage, but nevertheless shares the presence of the Lotus domain with other Tudor-family proteins (see Introduction). The *R. philippinarum* pooled transcriptome showed 5 transcript isoforms (1 did not cover the whole ORF but represented a truncated 5' transcript rather than a length isoform) that positively aligned with high quality values against *M. yessoensis* TDRD7A-like (best TBLASTX isoform hit: bitscore 288; e-value 1.72e-81; identity percentage 33.555%). These transcripts had almost identical sequences between each other and we considered them as between-individual polymorphisms (16 specimens were pooled to build the reference transcriptome), rather than actual biological isoforms: three transcripts (comprising the truncated one) had 100% amino acid identity and their sequence was uploaded on GenBank (accession code: MW170385); one had a single amino acid substitution; and one had an insertion of a single amino acid. Therefore, we could assess the presence of the TDRD7 orthologue (confirmed by a reverse BLAST against *M. yessoensis* that obtained as best hit the starting protein) in single copy in the *R. philippinarum* transcriptome. Moreover, the Sanger sequencing of 9 overlapping regions that comprehended the whole ORF confirmed the presence of the whole coding region of the transcript *in vivo*.

The transcript included an ORF of 1,134 amino acids (predicted molecular weight: 126.8 kDa), with 5 annotated domains (Figure 1): two Lotus domains in amino acid positions 20-92 and 361-429; three consecutive Tudor domains in amino acid positions 447-563, 644-764, and 962-1078.

Tdrd7 upregulation in the gonads

Tdrd7 was differentially transcribed between tissues, with a moderate but highly significant upregulation in the gonads (~2.5 times more transcribed considering pooled sexes; p-value = $1.4*10^{-12}$). On the other hand, the conserved germline marker *vasph* was transcribed ~8.5 times more in gonads than in somatic tissues (p-value = $3.3*10^{-24}$; pooled sexes). The difference between *tdrd7* and *vasph* lied at the level of gonadic transcription, that was almost double for the latter (average of normalized counts: 225.8 for *tdrd7* and 515.2 for *vasph*), rather than somatic transcription (88.6 against 60.7, respectively).

Considering sexes separately, the gonad differential transcription significance with respect to somatic tissues held for both transcripts (always upregulated in gonads), but the intensity of the signal was stronger in females (*tdrd7* ~3.1 times more transcribed in gonads; *vasph* ~11 times more transcribed in gonads) than in males (*tdrd7* ~2.1 times more transcribed in gonads; *vasph* ~5.5 times more transcribed in gonads). However, such differences were mostly confined to the transcriptional level of *vasph*, that was almost 3 times more transcribed in female gonads than in male gonads. Indeed, the transcription of *tdrd7* was statistically equal between the sexes. Lastly, the transcriptional level of the two genes in the somatic tissues were low and equal between the two sexes.

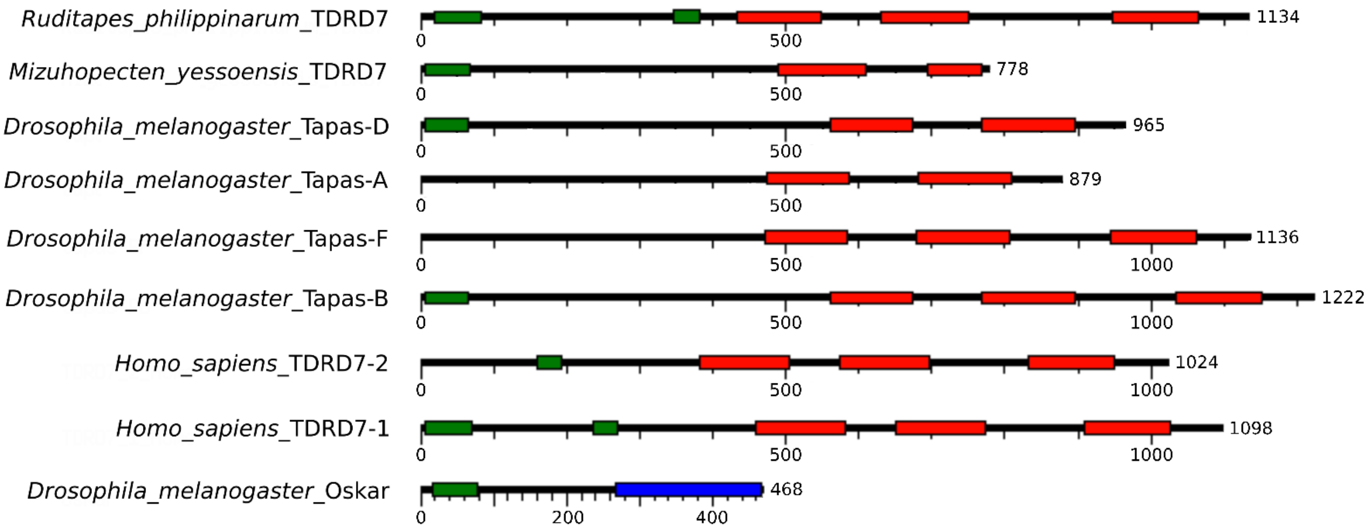

Ruditapes_philippinarum_TDRD7
1134
Mizuhopecten_yessoensis_TDRD7
778
Drosophila_melanogaster_Tapas-D
965
Drosophila_melanogaster_Tapas-A
879
Drosophila_melanogaster_Tapas-F
1136
Drosophila_melanogaster_Tapas-B
1222
Homo_sapiens_TDRD7-2
1024
Homo_sapiens_TDRD7-1
1098
Drosophila_melanogaster_Oskar
468
0
200
400
500
1000

Figure 1. Domain composition of TDRD7. Homologues in different species are depicted, together with *Drosophila* Oskar. Lotus (green) and Tudor (red) domains are highlighted. Oskar domain of Oskar is depicted in blue. For *D. melanogaster* and *H. sapiens* isoforms are shown for comparison with *R. philippinarum* and *M. yessoensis* domain composition.

Western blot results for the different antibodies are not consistent

Western blot was performed on male and female gonad homogenates of *R. philippinarum* by using all four antibodies (Figure 2). This analysis was not performed on animals sampled in reproductive spent phase due to the absence of gonads in that stage of the reproductive cycle.

For the clam-specific antibodies, the blot profiles were not always concordant between sexes and individuals, and the number and weight of the bands were variable. Between the two, usually anti-AYD profiles displayed more bands than anti-EKF. One band that was present in many specimens with both antibodies was approximately 37 kDa (Figures 2A and 2B). Other bands of 30, 50, 60, and 70 kDa were displayed in different blot profiles. Bands more easily relatable to the predicted molecular weight of the TDRD7 protein (126.8kDa) were observed for both antibodies in both sexes, but they were not always present. To test the specificity of antisera, the clam-specific antibodies were preincubated with a 20-fold molar excess of the peptides against which they were produced. This step was performed to chelate by competition every antigenic site of the primary antibody. Both controls showed significant reduction of some band intensity.

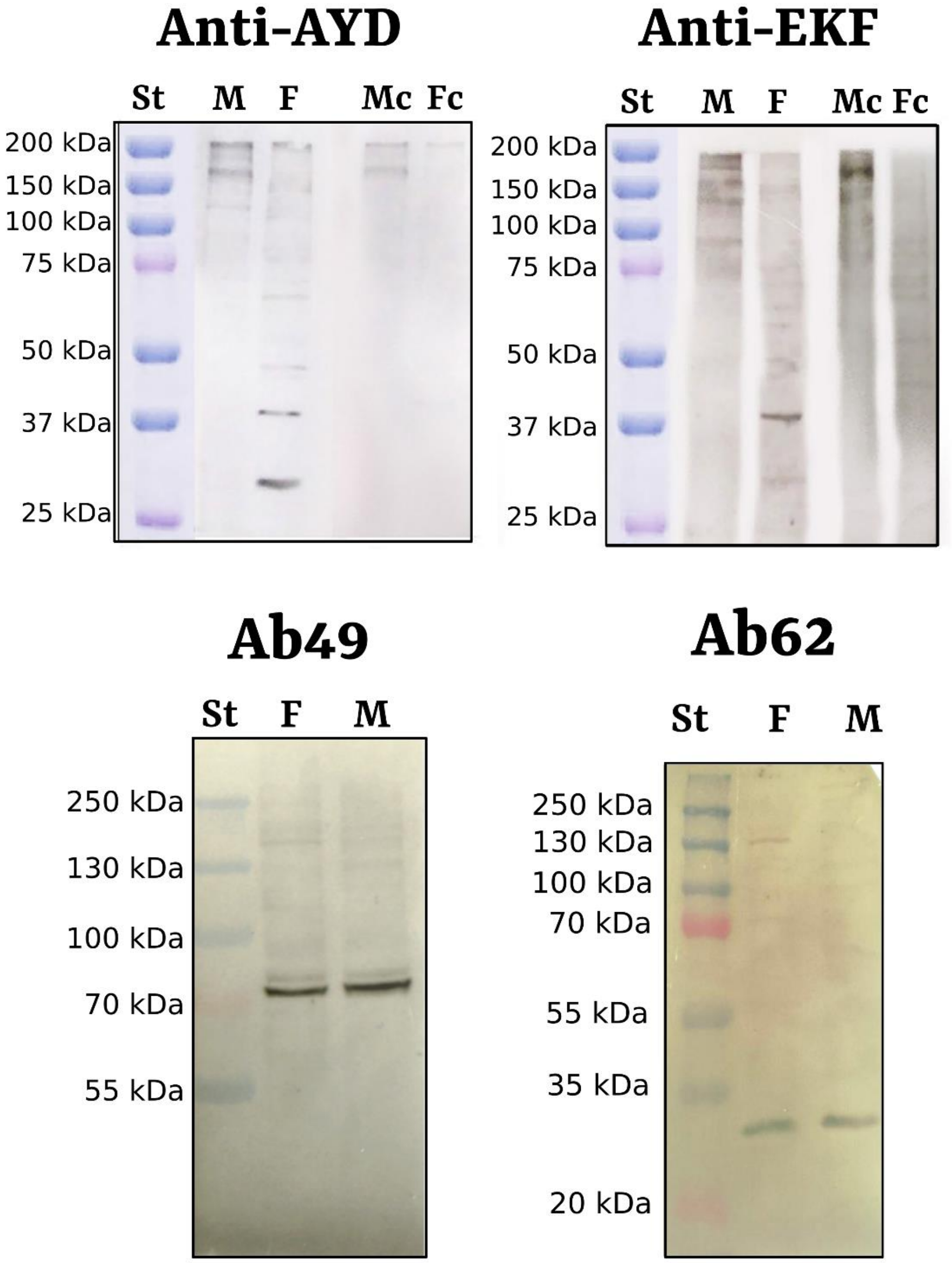


Figure 2. Western Blot analyses of the four anti-TDRD7 antibodies. Immunoblots were performed on testis (M) and ovary (F) extracts of specimens collected in the gametogenic phase. Standard molecular weights are depicted on the left of each blot. Anti-AYD and Anti-EKF (top blots) represent the two clam-specific antibodies, where clear and consistent bands were not present for the expected molecular weight. Mc and Fc represents immunoblots performed together with the control peptide. Ab49 and Ab62 (bottom blots) are the two commercial antibodies: in both cases a female-specific band of 150/160 kDa was present and consistent in the replicates. That band most likely represent TDRD7, since, despite the predicted molecular weight is ~127 kDa, the protein is known to appear heavier in SDS-PAGE-based blots (Hirose et al. 2000; Skorokhod et al. 2011).

More clear bands were present with commercial antibodies, however they did not exactly correspond to the predicted molecular weights (Figure 2C and 2D). Ab49, that was tested by the producer for WB, displayed two close bands in both sexes, one of which with the strongest staining, of ~80 kDa, therefore much lighter than the expected. Another very faint band was present in female samples only, at approximately 150/160 kDa. Ab62, on the other hand, did not display the 80 kDa bands, but a strong band of 30 kDa in both female and male samples, and a neat, again female-specific, band at ~150/160kDa. Other faint and scattered, probably non-specific, bands were present in both sexes.

Histological organization of the *Ruditapes philippinarum* gonad during the two stages of the reproductive cycle

During the gametogenic phase, in all sections, intestinal epithelium and acini full of developing gametes were present (for clear histological visualization of the investigated sites, refer to control IF samples at 40x magnification: Supplementary Figure C1). The intestinal epithelium was 100-200 µm thick and it consisted in a columnar single-cell layer lying on a basal lamina. Beneath it, connective tissue surrounded the gametogenic acini that constitute the gonads and separated them from the basal lamina.

In female samples, acini appeared full of oocytes, with clearly visible large nuclei. Germ cells in different meiotic phases were distinguished by size and by chromatin morphology through fluorescent nuclear staining. We refer to the oocyte maturation stages as described by Reunov and colleagues (2019): they define maturation stages of female germ cells from oogonia to mature oocytes. Their division was based on size and presence of nucleolus, but mostly to cytoplasmic germ granules that they observed through TEM. In the present study we mainly referred to size and the presence of nucleoli: small oogonia of ~11 µm diameter with nucleolus are present in the acinus periphery; primary oocytes and growing oocytes up to stage II lack nucleolus and range from 12 to 18 µm in diameter; growing oocytes of stage III are ~27 µm large and present a nucleolus; maturing oocytes represent the last stage before fertilization and their diameter is averagely 65 µm (Reunov et al. 2019).

Male acini were in general more compact and closer together compared to female acini. There was a clear centripetal organization of spermatogenesis within each acinus. From the periphery to the acinus lumen, spermatocytes, spermatids, and spermatozoa were recognizable by the different nuclear staining concentration and shape. The round and scarcely condensed nucleus of spermatocytes became tightly compressed and elongated in shape in spermatozoa.

Some samples were collected in the winter, during the spent phase, and in their sections no acini with gametes were present, as expected (see Supplementary Figure C2).

TDRD7 detection by immunofluorescent (IF) assay

Immunofluorescence images at the confocal microscope helped to appreciate the localisation of TDRD7 within single cells with respect to IHC thanks to higher resolution and contemporary nuclear staining. For IF images: all anti-TDRD7 antibodies are depicted in red, while dsDNA staining with TO-PRO3 dye is depicted in green.

- **Clam-specific antibodies**: anti-AYD and anti-EKF displayed similar marking patterns, therefore they will be considered together. Many anti-TDRD7 labelled cells were localised within the intestinal epithelium and were significantly different in shape from unstained columnar, batiprismatic cells (Figure 3A-B): labelled cells had a round nucleus and were often positioned close to the basal lamina. Most of these cells displayed anti-TDRD7 labelling within the nucleus (overlapping with TO-PRO3 was shown as yellow fluorescence in the images with merged channels). Some of these cells were clustered in small groups of 2-3 close to each other and were strongly and uniformly stained with anti-TDRD7 antibodies (Figure 3A). For some of these clusters it was possible to distinguish differently marked cells: one cell appeared exclusively labelled with anti-TDRD7 in its cytoplasm and no anti-TDRD7 staining was visible in the nucleus; the remaining cells, additionally to marked anti-TDRD7 cytoplasm, were stained also in the nucleus. This was observed in both male and female samples.

  Around both female and male acini, some cells were slightly stained by anti-TDRD7 (Figure 3C-D). However, differentiating germ cells within the acini completely lacked any anti-TDRD7 staining: oocytes at any maturation stage within the acini, just as spermatocytes, spermatids, or spermatozoa, were not marked.

  No evident anti-TDRD7 labelling was detected in the cells in the intestinal epithelium of adults sampled during the spent phase of reproductive cycle, except for some faintly marked cells in the intestine (Supplementary Figure C2A). Controls on male and female sections treated exclusively with secondary antibody showed very faint diffused labelling in all histological structure, but no cells showed higher fluorescence than the surroundings (Supplementary Figure C1A-B).
- **Commercial antibodies**: the two antibodies developed against human TDRD7 displayed different marking patterns. Ab49 marked exclusively germ cells of both sexes, and no specific labelling was observed in cells inside the intestine. In females, there was a strong cortical labelling in maturing oocytes (Figure 4A). In males, cells at the periphery of the acini (spermatogonia and spermatocytes) were uniformly tagged in the cytoplasm (Figure 4B). No labelling was present in spermatids and spermatozoa.

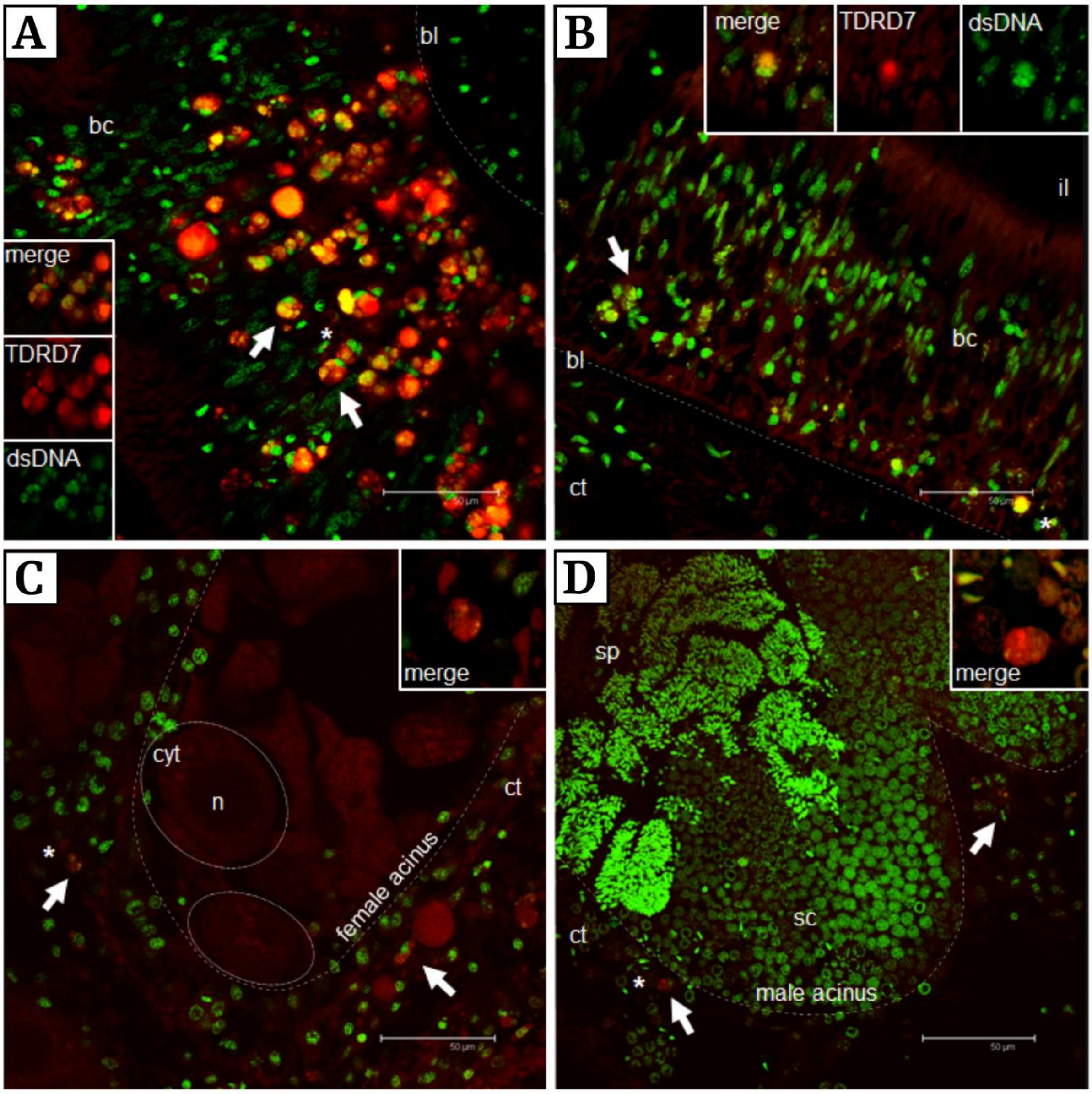


Figure 3. IF-localisation of clam-specific antibodies in gametogenic females and males. Here Anti-EKF is shown. **A**: marked female intestinal cells (arrows). Many cells are stained within the nucleus (yellow colour). The asterisk tags the section present in the inset on the left. **B**: marked male intestinal cells. The asterisk tags the section present in the inset on the top. **C**: female acinus and surrounding connective tissue. Two oocytes are highlighted within the acinus. Only some cells were slightly marked outside the acinus (arrow; asterisk tags the magnification in the top inset). **D**: male acini and surrounding connective tissue. Also in this case, few cells are slightly marked outside the acini (arrows; asterisk tags the magnification in the top inset). bc = batiprismatic cells; bl = basal lamina; ct = connective tissue; cyt = cytoplasm (of oocyte); n = nucleus (of oocyte); sc = spermatocytes; sp = spermatozoa. Red: anti-TDRD7 staining; Green: TO-PRO3 dsDNA dye.

On the other hand, with Ab62 we could observed tagged cells with round nuclei in the intestine. These cells were present in both sexes and presented a granular cytoplasmic anti-TDRD7 profile. Sometimes many of these cells were closely spaced in the intestinal section (Figure 5A), sometimes they were isolated (Figure 5B). No staining was observed in the nuclei. Female germ cells displayed some anti-TDRD7 labelling that was not uniformly

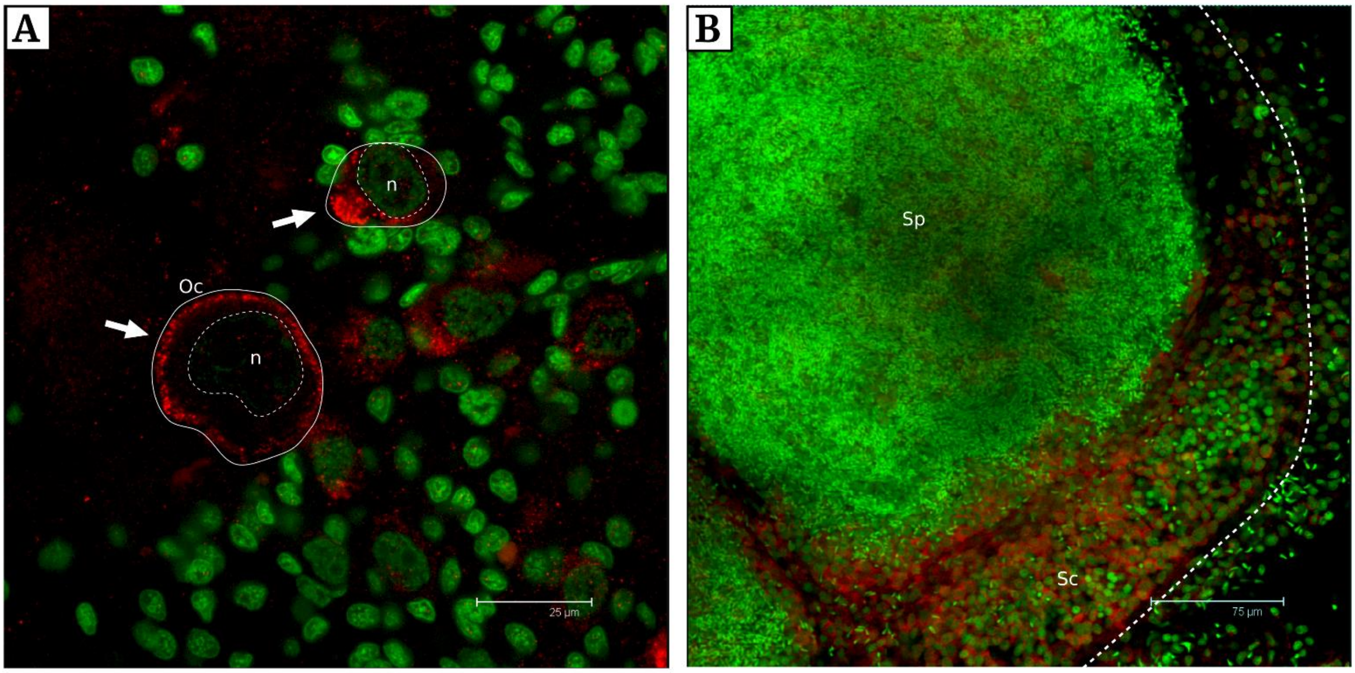


Figure 4. IF-localisation of Ab49 in gametogenic females and males. A: oocytes with cortical staining (arrows). **B**: male acinus with marked peripheral early differentiating germ cells. Oc = oocyte; n = nucleus (of oocyte); Sc = spermatocytes; Sp = spermatozoa. Red: Anti-TDRD7 staining; Green: TO-PRO3 dsDNA dye.

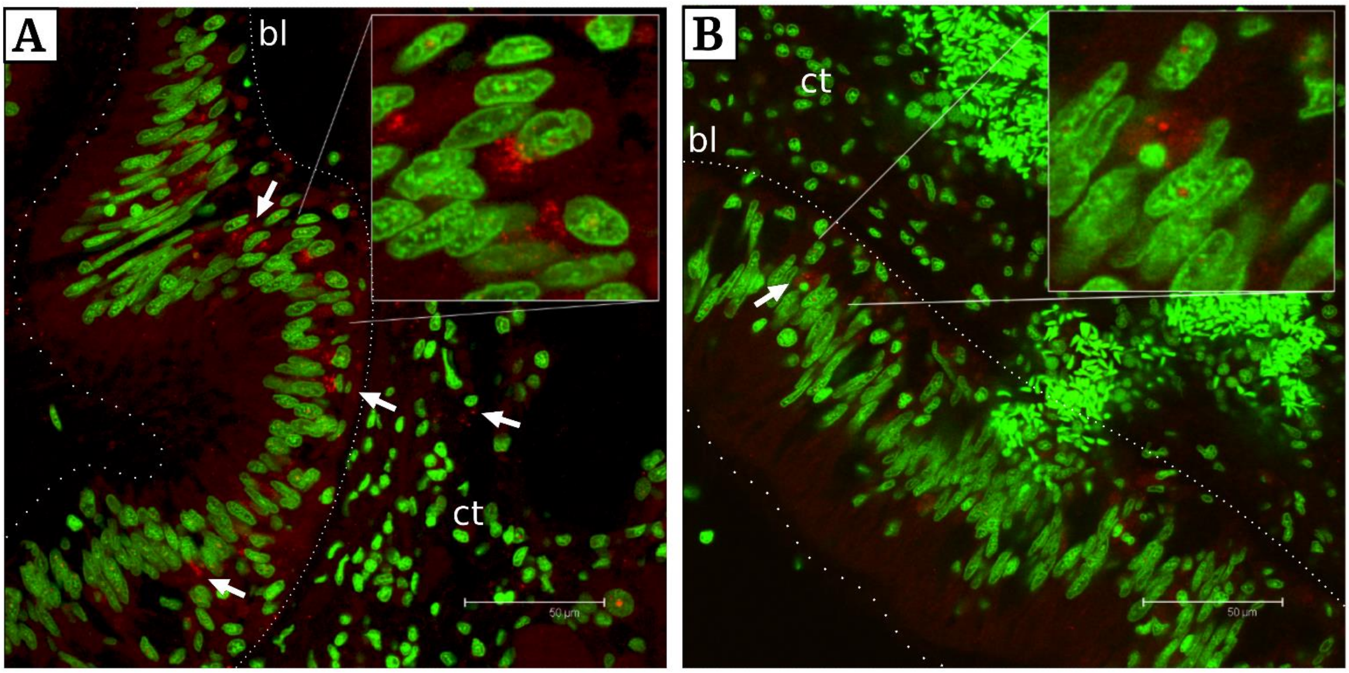


Figure 5. IF-localisation of Ab62 in the intestine of gametogenic females and males. A: female sample; **B**: male sample. Intestinal cells close to the basal lamina are marked with granular structures in the cytoplasm (arrows; examples magnified in insets). These cells differ from batiprismatic ones in having a round nucleus instead of an elongated flattened one. A marked cell in the connective tissue is highlighted on the right with an arrow. bl = basal lamina; ct = connective tissue. Red: Anti-TDRD7 staining; Green: TO-PRO3 dsDNA dye.

diffused in the cells, but rather organized in defined cytoplasmic regions with granular shape (that had different degrees of compactness). We could observe such cytoplasmic granules both in small germ cells of ~20 µm of diameter with nucleoli (Figure 6A-B), and in large maturing/mature oocytes (50 to 80 µm of diameter; Figure 6C-D). In male gonads, anti-TDRD7 was present in cytoplasmic granules in germ cells at early differentiation stages in the periphery of the acini (Figure 6E-F). Staining was never observed in the nuclei.

Specimens sampled during the reproductive spent phase displayed few cells labelled with Ab62 (Figure 7A-B). Specifically, cells with round nuclei and uniformly stained cytoplasm could be observed in the connective tissue adjacent to the intestinal basal lamina (Figure 7A), in the intestinal epithelium (Figure 7A), and dispersed in the connective tissue (Figure 7B).

Control sections without primary antibodies lacked any of the aforementioned labelling typical of both antibodies (Supplementary Figure C1C-D-E-F).

TDRD7 detection by immunohistochemistry (IHC) assay

- **Clam-specific antibodies:** During the gametogenic phase, in the intestinal epithelium both clam-specific TDRD7 antibodies stained only some cells located near the basal lamina and dispersed between the unlabelled batiprismatic cells that form the intestinal epithelium (Figure 8). The immunostaining was concentrated in almond-shaped structures of about 20 µm in length or smaller. No TDRD7-staining was visible in either oocytes or male germ cells, while slight staining was present in cells in the connective tissue close to the intestinal epithelium and morphologically similar to those in the intestine (Figure 8).
- In specimens in the spent phase, anti-TDRD7 staining was very faint in the same type of cells, in the intestinal epithelium and in the connective tissue (Supplementary Figure C2A). Negative control sections, in which the primary antibody was omitted, showed no staining over all histological structures (Supplementary Figure C3A-C).
- **Commercial antibodies:** Ab49 staining profile was consistent with the results obtained with IF. In female samples, the antibody strongly marked granular substance in the cortical region of all oocytes (Figure 9A). In male samples, spermatocytes in acinus peripheries were strongly marked, but a slight staining was observed also in the inner parts of the acini, where spermatozoa are present (Figure 9B). In both sexes, moreover, some cells in the connective tissue were stained (more numerous in the males), and the intestine was apparently non-specifically marked: in males some cells with round nuclei were present, but also batiprismatic cells were marked (Figure 9D); in females, a continuous marking pattern close to the basal lamina was present throughout the whole length of the intestine (Figure 9C).

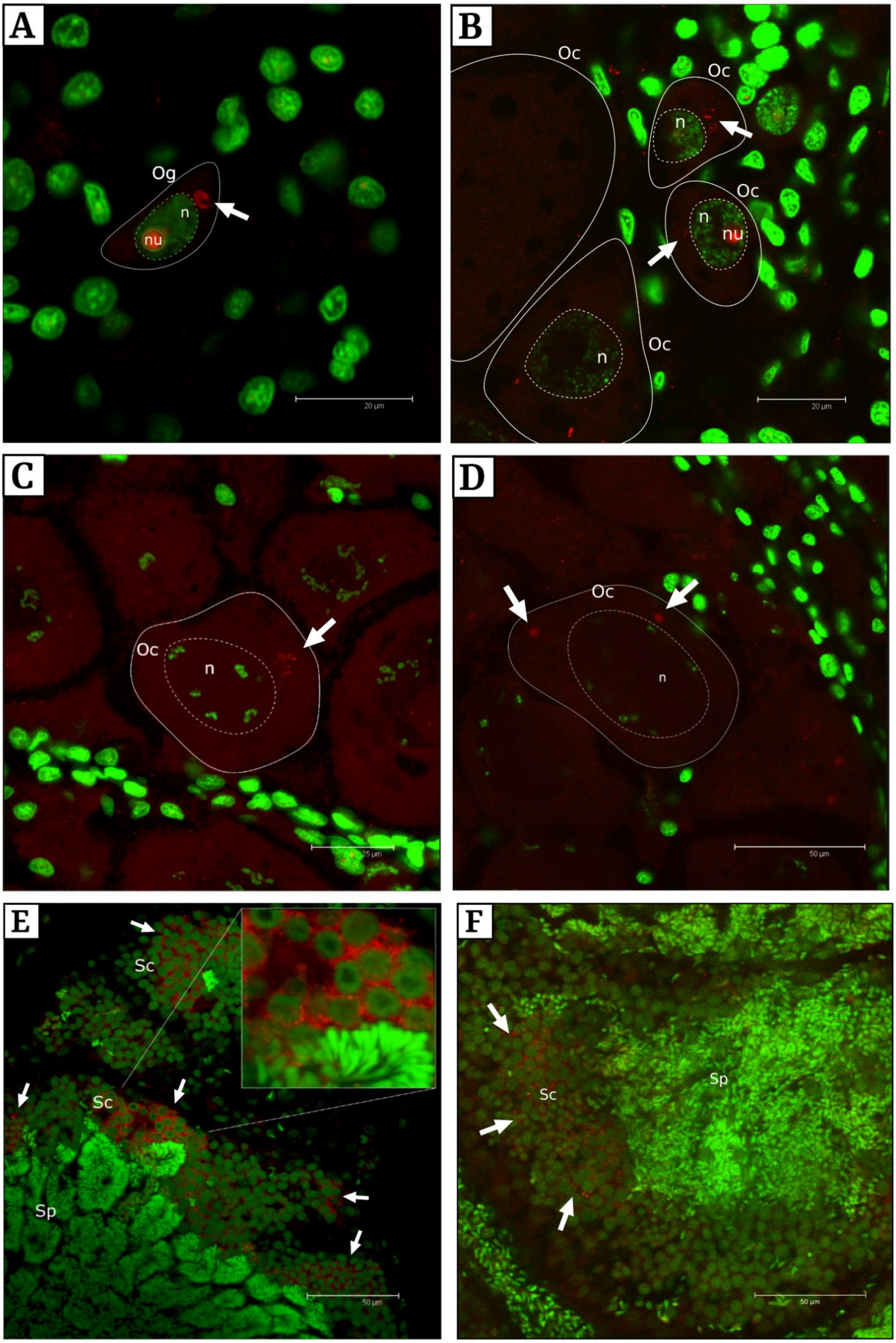

A
Og
n
nu
20 µm
B
Oc
Oc
n
Oc
n
nu
Oc
n
20 µm
C
Oc
n
25 µm
D
Oc
n
50 µm
E
Sc
Sc
Sp
50 µm
F
Sc
Sp
50 µm

Figure 6. IF-localisation of Ab62 in germ cells of female and male samples. (Figure in previous page) **A**: putative oogonium (or early growing oocyte) with nucleolus and evident anti-TDRD7 stained granule adjacent to the nucleus (arrow). **B**: early growing oocytes with nucleoli and anti-TDRD7 stained granular substance close to the nuclear envelope (arrows). **C-D**: mature oocytes with highly condensed chromatin and anti-TDRD7 stained granules in proximity of the nucleus (arrows). **E-F**: male acini with germ cells at diverse differentiation stages. Only peripheral cells, spermatocytes, are tagged with anti-TDRD7 (arrows). Oc = oocyte; Og = oogonium; n = nucleus; nu = nucleolus; Sc = spermatocytes; Sp = spermatozoa. Red: Anti-TDRD7 staining; Green: TO-PRO3 dsDNA dye.

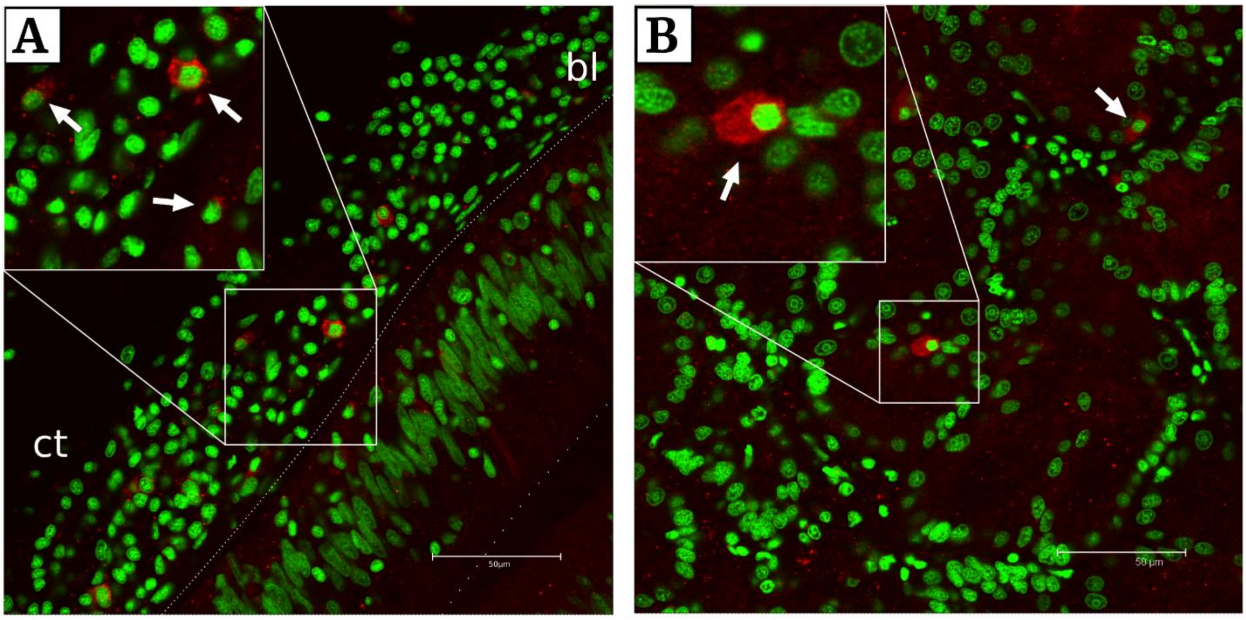


Figure 7. IF-localisation of Ab62 in connective and intestinal tissue during reproductive spent phase. **A**: Specimen of unknown sex showing anti-TDRD7 labelling in cells with round nuclei in the connective tissue in proximity to the intestinal basal lamina (white arrow). Also few cells within the intestinal epithelium are marked (one shown here in the magnification inset together with connective tissue ones). **B**: Marked cells (white arrows) are observable in the connective tissue of a specimen of unknown sex. ct = connective tissue; bl = basal lamina. Red: Anti-TDRD7 staining; Green: TO-PRO3 dsDNA dye.

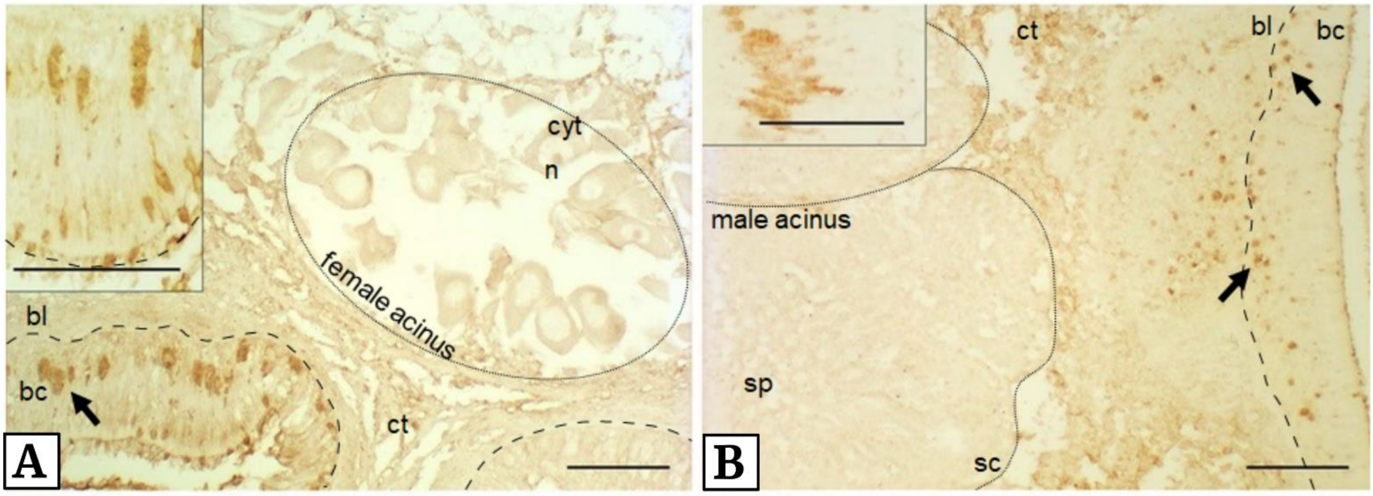


Figure 8. IHC-localisation of clam-specific antibodies in gametogenic females and males. Here, Anti-EKF is shown. **A**: female sample with marked cells in the intestinal epithelium (magnified in inset). **B**: male sample; also in this case only intestinal cells are marked (magnified in inset). bc = batiprismatic cells; bl = basal lamina; cyt = cytoplasm (of oocyte); ct = connective tissue; n = nucleus (of oocyte); sc = spermatocytes; sp = spermatozoa. Brown: Anti-TDRD7 staining. Scale bars = 100 μm.

Ab62 strongly tagged some cells with round nuclei in the intestinal epithelium close to the basal lamina in both sexes (Figure 10C-D). In females, anti-TDRD7 staining was observed also in maturing/mature oocytes (Figure 10A): some oocytes were stained uniformly, some displayed the antibody more concentrated at one side of the cytoplasm, and some others displayed small, marked granules in the cytoplasm (magnification of Figure 10A). Male acini, on the other hand, were only slightly labelled in peripheral cells (early differentiating ones; Figure 10B-D), hardly distinguishable from the surroundings, but clearly from the inner part of the acini. In both sexes, some cells in the connective tissue were marked (more numerous in males; Figure 10A-B-D). Negative control sections without primary antibodies showed no staining everywhere (Supplementary Figure C3B-D)

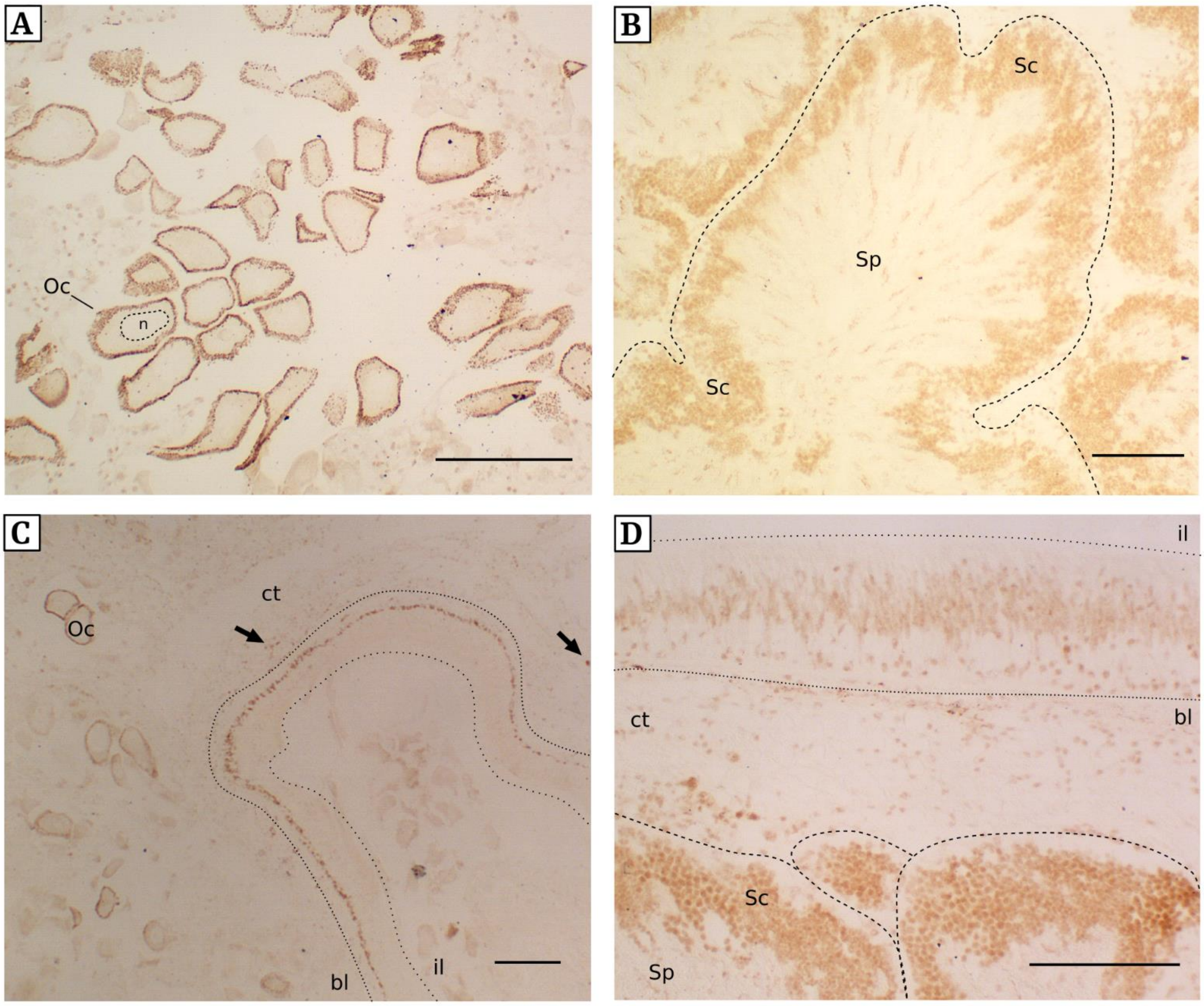


Figure 9. IHC-localisation of Ab49 in gametogenic females and males. (Figure in previous page) **A**: female sample with evident cortical staining in oocyte of various size (one is labelled in the figure and its nucleus is highlighted with dashed line). **B**: male acini with peripheral staining. Peripheral spermatocytes are highly stained, but some labelling is observable also in the inner part of the acini (one acinus is highlighted with dashed line). **C**: female sample with intestinal staining: the whole length of the intestine is marked. Some marked cells in the connective tissue are present (arrows). **D**: male sample with marked cells in the intestine and in the connective tissue. Some marked cells are close to the basal lamina, but batiprismatic cells are also tagged. bl = basal lamina; ct = connective tissue; il = intestinal lumen; Oc = oocyte; n = nucleus (of oocyte); Sc = spermatocytes; Sp = spermatozoa. Brown: Anti-TDRD7 staining. Scale bars = 100 µm.

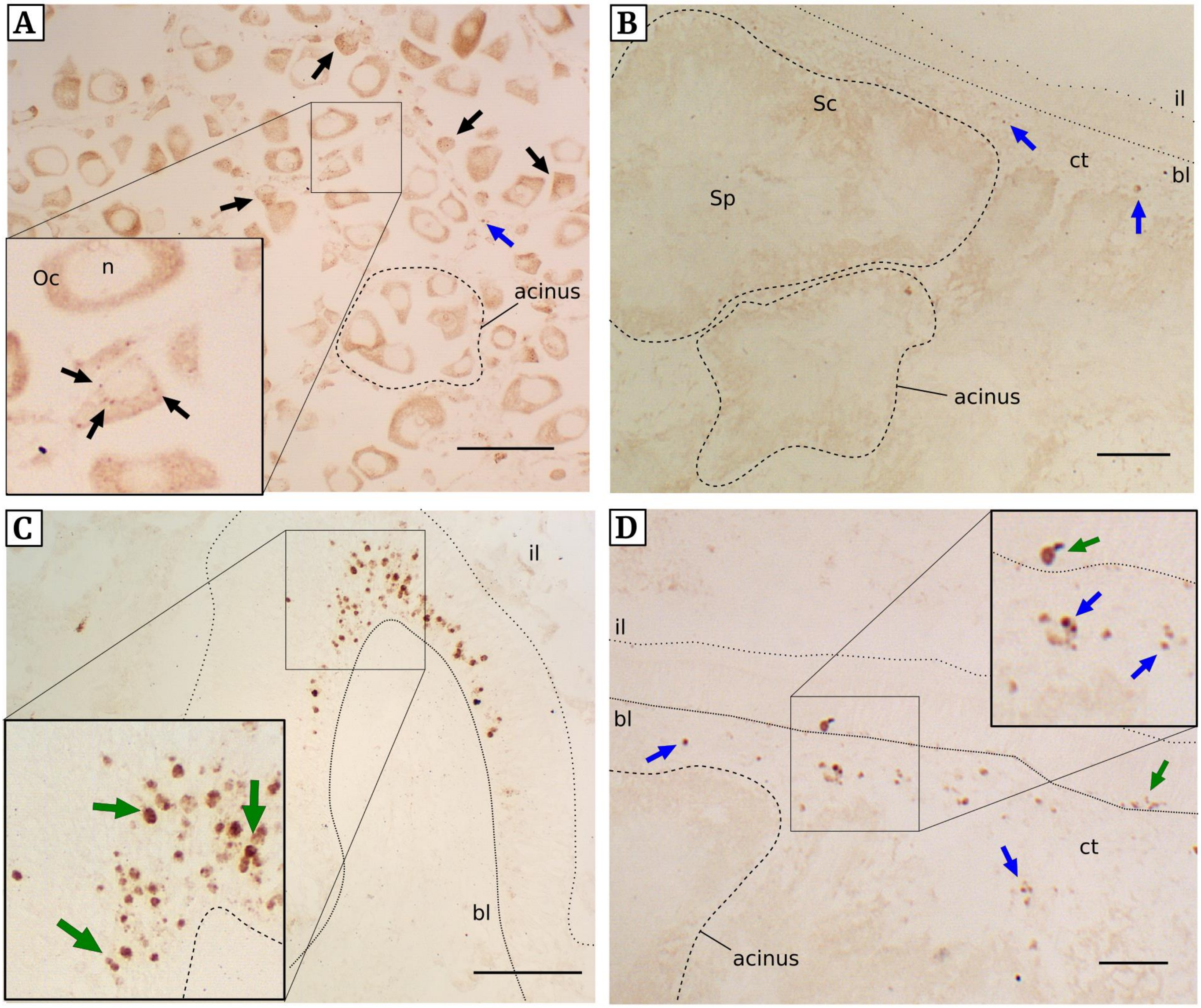


Figure 10. IHC-localisation of Ab62 in gametogenic females and males. **A**: female oocytes. The cytoplasm is uniformly stained but in several oocytes it is possible to observe condensed brown granules (black arrows; magnification in inset). Some tagged cells in layers of connective tissue between acini are present (blue arrow). **B**: male acini close to the intestine. Faint labelling is observable at the periphery of the acini, where early differentiating cells are present. Some tagged cells are present in the connective tissue (blue arrows). **C**: marked cells in the female intestinal epithelium (green arrows). **D**: marked cells in male connective tissue (blue arrows) and intestinal epithelium (green arrows). bl = basal lamina; ct = connective tissue; il = intestinal lumen; Oc = oocyte; n = nucleus; Sc = spermatocytes; Sp = spermatozoa. Brown: Anti-TDRD7 staining. Scale bars = 100 µm.

Discussion

A candidate protein involved in germline differentiation

In this work, we tried to identify factors that could act in germline specification and/or differentiation in *R. philippinarum*, that can possibly share functions as assemblers of germ granules with Oskar of *D. melanogaster*. It has been observed how, in *R. philippinarum*, some germ plasm-related structures are present in germ cells at initial stages of differentiation (Reunov et al. 2019). Since these germ granules are present at early germline stages in both sexes, the presence of a scaffolding protein, if not a germ plasm master regulator, seemed plausible in clams and maybe in bivalves in general. This putative protein, or proteins, would be able to establish germline fate in male and female annual gonad formation and possibly recruit Vasph, either directly or indirectly through germ plasm assembly. Recent studies on Oskar functional domains provided new models by which Oskar could promote germ plasm assembly by interaction between its Lotus domains and Vasa (Jeske et al. 2015, 2017). For this reason, we focused on the presence of Oskar-like proteins, or proteins containing homologous domains, in *R. philippinarum* transcriptome.

We performed a BLAST search of Oskar against publicly available sequences of bivalves, finding no orthologues sequences, as expected. However, the only hit we found was a TDRD7 homologue of *M. yessoensis*. The alignment similarity was confined to the Lotus domain that this protein shares with Oskar. We then annotated through BLAST the TDRD7 orthologue in our *R. philippinarum* transcriptome and observed that the gene was differentially transcribed in gonads (up to 3.1 times more transcribed with respect to somatic tissues in females).

In Metazoa, TDRD7 orthologues show conserved structural organization and present both Tudor and Lotus domains. With InterProScan we inferred the domain composition of *R. philippinarum* TDRD7 (Figure 1): two Lotus domains were present toward the N-t region of the protein and three consecutive Tudor domains were present toward the C-t region. The Tudor domain is commonly found in a wide range of proteins that are involved in RNA metabolism and splicing, histone modification, DNA damage response, cell division, differentiation, genome stability and gametogenesis (see Chapter B). Some of these proteins are metazoan innovations and are strictly associated to germline-related functions, and mostly in the piRNA pathway (see Chapter B). Many Tudor proteins of this kind display multiple Tudor domain that are thought to act as scaffolds to recruit Piwi and closely related piRNA pathway proteins (see Table 1 of Chapter B). Together with other GMP proteins (see Chapter A), these determinants are usually organized in ribonucleoprotein complexes of granular shape in germ cells (germ plasm-related structures) and are usually localised near nuclear pores (for instance, the intermitochondrial cement of spermatocytes and the chromatoid bodies of spermatids in *Drosophila* and mouse; Chuma et al. 2006; Handler et al. 2011; Yabuta et al. 2011).

The Lotus domain, on the other hand, is present in few proteins, the mostly characterized ones being TDRD7, TDRD5, and Oskar. The Lotus domain of Oskar was evaluated to be involved in both dimerization and Vasa-binding (Jeske et al. 2015, 2017). On the other hand, the same authors assessed that the homologous Lotus domains of TDRD5-7 could not dimerize but were nonetheless able to bind Vasa (Jeske et al. 2015, 2017). Interestingly, it has been observed how homologues of TDRD5 and TDRD7 are key-factors for the biogenesis and assembly of germ plasm-related structures of different species (mouse, fruit fly, and zebrafish), them being disorganized in their absence (Strasser et al. 2008; Tanaka et al. 2011; Yabuta et al. 2011; Patil et al. 2014; D'Orazio et al. 2020). The concerted functions of their domains, i.e. protein recruitment (Tudor domain; see Chapter B), and Vasa-interaction (Lotus domain; Hosokawa et al. 2007; Patil et al. 2014; Jeske et al. 2015, 2017) make them interesting investigational units for the characterization of germline and germline-related cytoplasmic supramolecular structures, in whose assembly they can be possibly involved. All these data allowed us to consider the TDRD7 orthologue as a worthy investigative unit for the study of germline specification and germ plasm assembly in *R. philippinarum*.

Four antibodies, some different results: which one to trust?

The four tested antibodies displayed different anti-TDRD7 marking profiles in *R. philippinarum* gonad tissues. Given that IF and IHC patterns did not always coincide for the same antibody, we decided to adapt a conservative approach and consider for each antibody exclusively those histological sites in which the two assays agreed. To summarize:

- Anti-EKF (clam-specific antibody built on the 1st Lotus domain) and anti-AYD (clam-specific antibody built on the 2nd Tudor domain) displayed ambiguous WB profiles, that were not consistent among samples and in which multiple bands were present. Bands of the predicted molecular weight of TDRD7 were sometimes slightly present, but not consistently across samples, and they never represented the most strongly stained ones. The histological profiles, on the other hand were curiously consistent among the two antibodies. However, properly defined germ cells were not labelled, and the staining profiles were confined to slightly marked cells in the connective tissue right outside the periphery of acini in both sexes. The strongest signal was observed in cells located within the intestinal epithelium. These cells had round nuclei (different from the flatted ones of the batiprismatic intestinal cells) and the antibody was often located in co-occurrence with dsDNA, i.e. within the nucleus.
- Ab49, commercial antibody that should bind the C-t edge of TDRD7, marked two bands of approximately 80 kDa in both sexes, and a light female-specific band of 150/160 kDa. In both IF and IHC, a strong germ cell-specific labelling was present for both sexes. The antibody localised in maturing/mature oocytes in an evident cortical position, while it localised in early

differentiating cells at the periphery of male acini. In both immunological assays, some cells of the connective tissue were stained, and non-specific marking of intestinal cells was present.

- Ab62, the other commercial antibody, that should bind a ~200 amino acid inter-domain region between the 2nd and 3rd Tudor domains (partially overlapping them both), marked a strong band of 30 kDa in both female and male samples and a single female-specific WB band of 150/160 kDa (corresponding to the one obtained with Ab49). IF and IHC assays coincided in marking some round-nucleus cells in the intestinal epithelium close to the basal lamina, but staining in the connective tissue was present mostly with IHC. Moreover, both immunological strategies marked female germ cells, where granular cytoplasmic structures usually close to the nuclear membrane were observable. The fact that, with IHC, granules were visible in only some oocytes was probably due to the fact that the IHC assay was performed on thin ~5 µm sections: given that mature oocytes can reach up 80 µm in diameter, the lack of granules in most cells was likely due to their absence in that specific cut section (on the other hand, IF was performed in ~150 µm sections, and with confocal microscopy it was possible to observe whole oocytes and acquire the optical section containing the granules). In male samples, with both IF and IHC, early differentiating germ cells at the acinus peripheries were observable, even if with the latter method the staining was very faint.

Given the different profiles for immunological assays and WBs of the different antibodies, it was clear that the *R. philippinarum* TDRD7 target was unlikely retrieved with all of them. The ambiguity of the WB profiles for the clam-specific antibodies casted some doubts on their specificity. Indeed, the lack of WB consistency across samples belonging to the same sex at the same developmental stage, and the fact that only intestinal cells were marked, allow us to consider these antibodies as probably non-specific for TDRD7. Moreover, TDRD7 was expected to be present in germline-related cells, and no properly defined germ cells were confidently marked by any of the two antibodies with any of the two immunological methods. It is true that what is expected based on other (model) organism observations not necessarily would meet confirmation in clams. It is true that some intestinal cells are thought to be involved in the annual renewal of gonads, as first suggested by the localisation of the Vasph protein in some round-nucleus cells close to the basal lamina (see Discussion below; Milani et al. 2015, 2017, 2018). However, TDRD7 is expected to be also involved in the proper assembly of granular cytoplasmic structure, and to interact with Vasph through its Lotus domain (see Introduction and Discussion below). The fact that most of clam-specific antibodies were located uniformly in the nucleus further lower the possibility that the actual target was TDRD7, given that Vasph was previously observed exclusively in the cytoplasm as regards intestinal cells (Milani et al. 2015, 2017, 2018). Skorokhod and colleagues (2011) observed the presence of a human 60 kDa isoform of TDRD7 that specifically localised in the nuclear fractions. This isoform lacked all Lotus and the 3rd Tudor domain, therefore is probably not involved in Vasa-binding functions (given that it is the Lotus domain to cover that activity; Jeske et al. 2017) but could cover DNA-related activities (like many other Tudor proteins do: see Chapter B). The light WB bands that we observed for the clam-specific antibodies might represent putative TDRD7 isoforms, some of which could be localised in nuclei (like how observed in the intestinal cells). However transcriptomic data did not show any alternative spliced transcripts and, while the possibility cannot be completely excluded, this would not explain the lack of the expected full-protein immunological profiles. To conclude, we believe that the clam-specific antibodies unlikely target TDRD7.

To discriminate the confidence of Ab49 and Ab62 in TDRD7 targeting we could rely both on WB profiles and on expression pattern. Both antibodies targeted a 150/160 kDa band in female samples, but Ab49 additionally targeted a couple of close bands around 80 kDa in both sexes, one of which showing the strongest labelling, and also Ab62 marked a lighter band in both sexes of a lower molecular weight, that is 30 kDa. Based on the amino acid sequence, the predicted molecular weight of *R. philippinarum* TDRD7 is 126.8 kDa, that is similar to the predicted weight for the human orthologue: 123.6 kDa. However, previous works in human cells assessed that in SDS-PAGE-based WBs the longest isoform corresponded to a 160 kDa band (Hirose et al. 2000; Skorokhod et al. 2011). Given the same domain composition, we believe that any reason that makes TDRD7 to run slower on an SDS-PAGE-base electrophoresis should be valid for both orthologues. For this reason, the 150/160 kDa band that we observed in female samples for Ab49 and Ab62 was indeed most likely TDRD7.

However, considering the labelling on tissues, Ab62 appeared to be the best candidate for specificity. This is supported by the fact that previous studies on *R. philippinarum* (Reunov et al. 2019) reported a clear expression pattern for Vasa that is actually similar to what we obtained with Ab62: given that Vasa and TDRD7 should interact (Jeske et al. 2015, 2017), a partially similar localisation was indeed a logical expectation. Instead, the cortical staining of oocytes observed with Ab49 was never observed for any Tudor protein, and it is more parsimonious to think that such staining is due to Ab49 higher specificity for some other protein. The 30 and 80 kDa bands might again be interpreted as putative isoforms, but no transcriptional evidence so far confirm it, given that any alternatively spliced isoform should have been present in the assembled transcriptome. However, it would be interesting to investigate the nature of the strongest band with lower weight, as, for example, through MALDI-MS for the identification of proteins by peptide mass fingerprinting.

To conclude, we believe that Ab62, for the presence of an expected band at 150/160 kDa and for the expression profile, specifically marks *R. philippinarum* TDRD7, allowing us to discuss its histological profiles.

TDRD7 is localised in putative undifferentiated germ cells in the intestinal epithelium

Through immunolocalisation experiments, female and male specimens of *R. philippinarum* were observed at two stages of development (gametogenic and spent phase) to identify the localisation of TDRD7 in tissues and cell types. Differences existed between the staining localisation in gametogenic males and females. However, TDRD7 was consistently localised in cells in the intestinal epithelium, forming small clusters of few, closely associated cells.

The annual renewal of gonads in clams is preceded by proliferation of Vasph-tagged cells within the intestinal epithelium (Milani et al. 2017, 2018). In the present study we could observe anti-TDRD7 staining in similar cells close to the basal lamina of the intestine. These cells had different nuclear shapes in respect to elongated cells of the intestinal epithelium, and we propose that they corresponded to the characterized Vasph-tagged cells observed in previous works (Milani et al. 2015, 2018). These intestinal clusters were interpreted as totipotent cells involved in the annual gonad renewal (we will refer to these cells as PriSCs to be in line with previous interpretation: whether they actually maintain somatic potential like intestinal regeneration, i.e. proper PriSCs, or only germ cell fate, that would make them PGCs, is still a matter of debate and more cytological analyses are needed). Indeed, these cells were observed in small numbers during the spent phase, and in greater numbers during the reproductive season. The interpretation for it was that these cells represent the winter repositories of PriSCs that, following inductive seasonal stimuli, start to divide at the beginning of the reproductive season. Then, they would migrate into the connective tissue and populate or form new acini, where they would start their actual germ cell-specific differentiation, eventually producing gametes.

Similar Vasa-tagged intestinal cells were observed also for other bivalves like *M. arenaria*, suggesting that the seasonal proliferation of germ cells might undergo similar mechanisms at least in some bivalve species (Milani et al. 2017). However, since the seasonality of the gonad ripening is shared by all bivalves, undifferentiated germ cells might localise in different positions in the different species, stimulating interest in such studies for a larger set of species, also considering the wide phylogenetic distances between families of Bivalvia. What we could observe in the present study was the fact that also TDRD7 is present in the clam intestinal clusters, further suggesting their characterization as either PriSCs or PGCs. Also stained cells in the connective tissue were present, even if much less frequently with the IF method than with IHC. Moreover, we observed marked cells in the connective and intestinal tissues also during the reproductive spent phase, coherently with their supposed role as germ cells repositories between annual reproductive cycles. In line with previous interpretations, these could be the putative migrating PGCs previously observed with Vasph, i.e. the "bridge" between the intestinal clusters and the acini (Milani et al. 2015, 2017, 2018).

TDRD7 is expressed at different stages of gamete differentiation

In the present study we also observed anti-TDRD7 staining in differentiating germ cells within acini of both sexes. However, while oocyte labelling was confirmed with the independent IHC method, male germ cells were only slightly stained with it. This, added to the fact that also with IF apparently not the totality of early male germ cells at the same differentiation stage were stained and to the fact that Ab62 marked an evident WB band of the expected TDRD7 weight exclusively in females, address us to be cautious in interpreting these results, and more assays are needed in male specimens to either confirm or revisit the pattern. On the other hand, female germ cells displayed consistent patterns: we could observe TDRD7 in granular structures in different oocyte differentiation stages. Most of the times these granules were observed in mature oocytes, but sometimes also in relatively small ones.

Reunov and colleagues (2019) defined different *R. philippinarum* germ cell maturation stages. Their work was mostly focused on the presence of germ granules observed with immuno-TEM during the different differentiation processes of the germ cells in clams. They observed the presence of cytoplasmic perinuclear granules tagged with Vasph in both male and female early differentiating germ cells, namely spermatogonia and oogonia, that were proposed as involved in the mitosis-to-meiosis shift in both sexes. Then the granules disperse and Vasph was observed as perinuclearly scattered with no evident granular structures. In female germ cells, however, granule assembly is resumed twice: once in the early stages of oocyte grow, and once in the late stages of oocyte maturation.

Following their observations, it is tempting to associate the TDRD7 granules that we observed to those characterized based on immuno-TEM. While mature ones are easily identifiable by the size and by the highly condensed chromatin typical of interrupted meiosis (Figure 6C-D), it is not straightforward to identify the differentiation stage of the smaller ones. These oocytes had a relatively uniform dsDNA staining, and they had a nucleolus, but their sizes varied in range. They might represent oogonia and/or growing oocytes in the first stages. For instance, the germ cell depicted in Figure 6A might be an oogonium, due to the presence of a nucleolus and a size that goes from 20 μm for the long axis, to 10 μm for the short one. Moreover, the chromatin staining is very diffused, and is coherent with a stage that precede meiosis onset. Those depicted in Figure 6B, on the other hand, could be growing oocytes at initial stages, given the much larger size (~20x30 μm), and a higher degree of chromatin compactness. Reunov and colleagues observed the complete lack of granules in growing oocytes at stage III, that are those where a nucleolus was present after being absent it in stage I and stage II. Cells in Figure 6B have a nucleolus, therefore they should be in stage III. However, in Reunov and colleagues work, optical microscopy pictures show the presence of nucleoli also in oocytes with sizes comparable to stages I and II, therefore it might mean that such feature is not straightforwardly associated to oocyte stage, or simply that size is not a strong indicator of it.

Unfortunately, the resolution of our analyses did not allow us to observe with high confidence chromatin structures like synaptonemal complexes that might have clarified the stage. However, oocytes depicted in Figure 6B represent nevertheless early oocytes, and it is tempting to associate the TDRD7 granules with the Vasph-tagged ones observed in the early oocyte differentiation stages. More observations comprehending earlier stages of gonad differentiation might clarify the pattern. Indeed, most of the oocytes in our sections were mature: specimens at the beginning of the reproductive season should have much more germ cells at earlier differentiation stages, and the ratio between oogonia and primary oocytes with respect to mature oocytes should be biased toward the former ones. This would allow us to observe a much higher number of early germ cells, that in the present study represented the minority.

The most widely distributed TDRD7 pattern in female germ cells was the presence of evident granules in large mature oocytes (Figure 6C-D). Reunov and colleagues (2019) observed the reappearance of Vasph-tagged germ granules in the late stages of oocyte maturation. Granules of earlier stages were associated to meiotic activities, as said before, but those of the mature stages were not, given that meiotic mechanisms should have already been interrupted, on hold for fertilization. They proposed that these late granules might be involved in preformation mechanisms of germ cell specification in the embryo, a mechanism that have been proposed for clams (Milani et al. 2015, 2018). These granules would be selectively inherited in embryo PriSCs, whose progeny in adult tissue could be represented, following the described hypothesis, by the intestinal GMP-tagged cells. It is again tempting to associate the late oocyte TDRD7 granules to those containing Vasph, closing the circle from intestinal PriSCs/PGCs, to early differentiating oocytes, mature oocytes, and eventually back to intestinal PriSCs/PGCs. These are still speculation following previous Vasph-based interpretations and additional analyses should be done on the matter: most of all, performing the contemporary staining of Vasph and TDRD7 to search for co-localisation of fluorescent signals (already planned by using combination of polyclonal anti-rabbit and monoclonal anti-mouse antibodies) and also performing similar immunological assay on embryos to confirm or confute the proposed germline specification mechanism. Moreover, further characterization by immuno-TEM analyses of TDRD7 would allow us to assess whether the granules that we observed with anti-TDRD7 actually coincide with those observed containing Vasph.

TDRD7 might be involved in the proper assembly of *R. philippinarum* germ granules

In all cases (from intestinal cells to early and late oocytes), TDRD7 was always marked in granular structures. As said before, the co-occurrence of multiple Lotus and Tudor domains makes TDRD7 a putative scaffolding protein able to recruit both Piwi and Vasa. Indeed, it has been already associated to proper germ granules formation in different animals. In mice, it has been observed how the protein is crucial for the biogenesis and assembly of male germline chromatoid bodies (Tanaka et al. 2011). Interestingly, similar results were obtained also for TDRD5, the other Lotus-Tudor-containing protein, that was assessed as fundamental for the proper assembly of the same cytoplasmic structures (Yabuta et al. 2011). Also in *Drosophila*, these two proteins have been observed in similar functions and districts: the insect Tejas (orthologue of TDRD5) and Tapas (orthologue of TDRD7) are crucial for the nuclear localisation of Piwi (Patil et al. 2014). Tapas interacts with Vasa and piRNA pathway proteins allowing their localisation in the nuage, a perinuclear supramolecular structure equivalent to mouse chromatoid bodies (Patil et al. 2014). Another model organism in which TDRD7 has been associated with proper formation of germ cell perinuclear granules is *D. rerio*. In zebrafish, disruption of granule architecture was observed after TDRD7 loss-of-function (Strasser et al. 2008), and mis-localisation of germ plasm due to TDRD7 disruption led to somatic differentiation of PGCs (D'Orazio et al. 2020) Moreover, TDRD7 is also involved in the formation of cytoplasmic structures in the mouse embryonic ocular lens formation, i.e. in somatic tissues (Lachke et al. 2011). There, the protein is involved in RNA-recruitment and the formation of RNA-granules (Lachke et al. 2011), suggesting that the control of ribonucleoprotein aggregates is the common mode of action of TDRD7 in different species and tissues.

In our study we observed TDRD7-tagged granular structures in putative PriSCs localised in the intestinal epithelium and the connective tissue, and in differentiating oocytes, curiously similarly to previous results with Vasph. The molecular activity of TDRD7 domains and its immunological patterns allow us to propose a physical interaction between it and Vasa, with putative roles in the germ granule assembly. However, so far, the fact that both the intestinal cells and the germ cell granules observed with the two different antibodies are equivalent, is still hypothetical, invoking the need for further analyses that would allow us to actually confirm the contemporary presence of the two proteins in the same structures.

Conclusions

With the present work, we suggest TUDOR domain-containing protein 7 (TDRD7) as a possible candidate acting in the assembly of *R. philippinarum* germ cell granules. This is supported by literature data on Lotus-Tudor-containing proteins (comprising TDRD7 homologues), as well as by *in situ* localisation of TDRD7 in putative germ cells, in both their undifferentiated stage in both sexes intestinal epithelium, and in granular structures of different oocyte maturation stages. Our interpretation is that TDRD7-immunolabeled cells within the intestinal epithelium might be Primordial Stem Cells (PriSCs), precursors of both Primordial Germ Cells (PGCs) and cells of the somatic lineage (Solana 2013). Our histological observations in cell populations previously documented as Vasph-tagged undifferentiated germ cells provide good evidence for it. Moreover, the identification in oocytes of TDRD7-stained granular structures in similar stages of previously observed Vasph-tagged germ granules (Reunov et al. 2019) led us to temptingly associate the two patterns to the same granular structures. However, to solidly validate such hypotheses, future analyses are needed, including MALDI-MS for the identification of protein WB bands, TDRD7-Vasph co-tagging and immunoprecipitation analyses to confirm the physical interaction between the two proteins, and immuno-TEM observations of TDRD7 granules to better characterize them.

Moreover, a final consideration is worth to be made on the nature of the target organism, since investigation on non-model animals is sewed with obstacles. The lack of high-quality genomes, of adjusted protocols, and of tested target-specific dyes and antibodies, higher the difficulties of exploration in such organisms. However, the value of non-model animal investigation is evident and crucial for a deep understanding of the biological processes. Indeed, comprehending their complexity and variability can be eased only with the aid of numerous and heterogeneous representatives, trying to approximate as much as possible the extant diversity that covers all aspects of life

Note:

The results exposed in the present chapter are currently being expanded and integrated, in sight of publication on a journal with IF in the near future.

Final considerations

Within the present PhD thesis, different aspects of germline differentiation were covered and investigated. Spacing from a metazoan-wide analysis of generic molecular pathways related to the germline, to the evolutionary pathways followed by a protein family in animal phyla, and to the species-specific characterization of a putative docking platform protein involved in germ granule assembly, the scopes and methods I used strongly differed. In each chapter, conclusions based on the obtained results were drawn, them standing on variable ground solidity. In taking the stock of the thesis, I do not intend to retrace what has already been said in the proper context throughout its development. What can be said, however, is that the different and diverse approaches that I followed embody the multifaceted biological question from which these projects stemmed. This underlines the need for the use of different approaches and for the appropriation of a diversified set of methods and skills.

I used bioinformatic tools for transcriptomic analyses and molecular sequence evolution, together with wet lab techniques, as, for example, *in situ* visualization of target proteins. The advantages of multiple approaches for investigating complex issues are evident. No single method, and no single point of view can uncover the tangled mechanisms under the determination and differentiation of the germline: if theoretical and experimental models are presently available is thanks to the effort of multiple disciplines and perspectives throughout many decades. Moreover, when considering the large evolutionary distances spanned by organisms that share the features considered, with all their distinctive declensions, the complexity arises even more. During my PhD experience, I tried to collect a diversified set of skills and to use different approaches to investigate the complexity of the topic by different sides. The results reported in the present thesis are not conclusive answers by themselves, and much more investigation is needed for each and every one of them. During the course of each chapter, pros and cons, together with flaws and qualities of each experimental asset and result were discussed, and perspectives for a future developing were addressed. Indeed, like any research, the end dresses in meaning only when it develops in the beginning of the subsequent step forward.

Bibliography

Abascal, F., Corpet, A., Gurard-Levin, Z. A., Juan, D., Ochsenbein, F., Rico, D., Valencia, A., & Almouzni, G. (2013). Subfunctionalization via adaptive evolution influenced by genomic context: The case of histone chaperones ASF1a and ASF1b. *Molecular Biology and Evolution*, 30(8), 1853–1866.

Aboobaker, A. A., & Blaxter, M. L. (2003). Hox Gene Loss during Dynamic Evolution of the Nematode Cluster. *Current Biology* 13, 37-40.

Alexa, A., Rahnenfuhrer, J. (2021).TopGO: Enrichment Analysis for Gene Ontology. R Package Version 2.46.0

Alié, A., Hayashi, T., Sugimura, I., Manuel, M., Sugano, W., Mano, A., Satoh, N., Agata, K., & Funayama, N. (2015). The ancestral gene repertoire of animal stem cells. *PNAS*, 112(51), E7093–E7100.

Anantharaman, V., Zhang, D., & Aravind, L. (2010). OST-HTH: a novel predicted RNA-binding domain. Biology Direct, 5(13).

Anisimova Editor, M. (n.d.). *Evolutionary Genomics Statistical and Computational Methods Second Edition Methods in Molecular Biology 1910*.

Anne, J. (2010). Arginine methylation of SmB is required for *Drosophila* germ cell development. *Development*, 137(17), 2819–2828.

Arora, K., & Corbett, K. D. (2019). The conserved XPF:ERCC1-like Zip2:Spo16 complex controls meiotic crossover formation through structure-specific DNA binding. *Nucleic Acids Research*, 47(5), 2365–2376.

Artigas, G. Q., Lapébie, P., Leclère, L., Takeda, N., Deguchi, R., Jé kely, G., Momose, T., & Houliston, E. (2018). A gonad-expressed opsin mediates light-induced spawning in the jellyfish *Clytia*. eLife, 7:e29555.

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., Williams, Z., Ohler, U., & Tuschl, T. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, 492(7429), 382–386.

Ayyanathan, K., Lechner, M. S., Bell, P., Maul, G. G., Schultz, D. C., Yamada, Y., Tanaka, K., Torigoe, K., & Rauscher, F. J. (2003). Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: A mammalian cell culture model of gene variegation. *Genes and Development*, 17(15), 1855–1869.

Balakrishnan, L., & Bambara, R. A. (2013). Flap endonuclease 1. *Annual Review of Biochemistry* 82, 119–138.

Bast, J., Schaefer, I., Schwander, T., Maraun, M., Scheu, S., & Kraaijeveld, K. (2016). No accumulation of transposable elements in asexual arthropods. *Molecular Biology and Evolution*, 33(3), 697–706.

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489.

Bauer, K. M., Dicovitsky, R., Pellegrini, M., Zhaxybayeva, O., & Ragusa, M. J. (2019). The structure of a highly-conserved picocyanobacterial protein reveals a Tudor domain with an RNA-binding function. *Journal of Biological Chemistry*, 294(39), 14333–14344.

Beltra, T., Barroso, C., Birkle, T. Y., Stevens, L., Schwartz, H.T., Sternberg, P. W., Fradin, H., Gunsalus, K., Piano, F., Sharma, G., Cerrato, C., Ahringer, J., Martinez-Perez, E., Blaxter, M., Sarkies, P. Comparative epigenomics reveals that RNA polymerase II pausing and chromatin domain organization control nematode piRNA biogenesis. *Dev Cell* 48, 793-810.

Bertocchini, F., & Chuva de Sousa Lopes, S. M. (2016). Germline development in amniotes: A paradigm shift in primordial germ cell specification. *BioEssays* 38(8), 791–800.

Bian, C., Xu, C., Ruan, J., Lee, K. K., Burke, T. L., Tempel, W., Barsyte, D., Li, J., Wu, M., Zhou, B. O., Fleharty, B. E., Paulson, A., Allali-Hassani, A., Zhou, J. Q., Mer, G., Grant, P. A., Workman, J. L., Zang, J., & Min, J. (2011). Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. *EMBO Journal*, 30(14), 2829–2842.

Björklund, Å. K., Ekman, D., & Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Computational Biology*, 2(8), 0959–0970.

Blaxter, M., & Koutsovoulos, G. (2015). The evolution of parasitism in Nematoda. *Parasitology*, 142, S26–S39.

Blondel, L., Jones, T. E. M., & Extavour, C. G. (2020). Bacterial contribution to genesis of the novel germ line determinant Oskar. *ELife*, 9, e45539.

Boke, E., Ruer, M., Wühr, M., Coughlin, M., Lemaitre, R., Gygi, S. P., Alberti, S., Drechsel, D., Hyman, A. A., & Mitchison, T. J. (2016). Amyloid-like Self-Assembly of a Cellular Compartment. *Cell*, 166(3), 637–650.

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

Bontems, F., Stein, A., Marlow, F., Lyautey, J., Gupta, T., Mullins, M. C., & Dosch, R. (2009). Bucky Ball Organizes Germ Plasm Assembly in Zebrafish. *Current Biology*, 19(5), 414–422.

Botuyan, M. V., & Mer, G. (2016). Tudor Domains as Methyl-Lysine and Methyl-Arginine Readers. In *Chromatin Signaling and Diseases* (curated by Binda O., Fernandez-Zapico, M. E.) Elsevier Inc, 2016, pp. 149–165.

Bourquin, J.-P., Stagljar, I., Meier, P., Moosmann, P., Silke, J., Baechi, T., Georgiev, O., & Schaffner, W. (1997). A serine/arginine-rich nuclear matrix cyclophilin interacts with the C-terminal domain of RNA polymerase II. *Nucleic Acids Research* 25(11), 2055–2061.

Breton, S., Beaupré, H. D., Stewart, D. T., Hoeh, W. R., & Blier, P. U. (2007). The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? In *Trends in Genetics* 23(9), 465–474.

Brunet, T., & King, N. (2017). The Origin of Animal Multicellularity and Cell Differentiation. *Developmental Cell* 43(2), 124–140.

Brunet, T., & King, N. (2020). The single-celled ancestors of animals: a history of hypotheses. *Preprints* 2020110302.

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368.

Bunting, S. F., Callén, E., Wong, N., Chen, H. T., Polato, F., Gunn, A., Bothmer, A., Feldhahn, N., Fernandez-Capetillo, O., Cao, L., Xu, X., Deng, C. X., Finkel, T., Nussenzweig, M., Stark, J. M., & Nussenzweig, A. (2010). 53BP1 inhibits homologous recombination in brca1-deficient cells by blocking resection of DNA breaks. *Cell*, 141(2), 243–254.

Callebaut, I., & Mornon, J. P. (2010). LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinformatics*, 26(9), 1140–1144.

Callen, E., di Virgilio, M., Kruhlak, M. J., Nieto-Soler, M., Wong, N., Chen, H. T., Faryabi, R. B., Polato, F., Santos, M., Starnes, L. M., Wesemann, D. R., Lee, J. E., Tubbs, A., Sleckman, B. P., Daniel, J. A., Ge, K., Alt, F. W., Fernandez-Capetillo, O., Nussenzweig, M. C., & Nussenzweig, A. (2013). 53BP1 mediates productive and mutagenic DNA repair through distinct phosphoprotein interactions. *Cell*, 153(6), 1266–1280.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10, 421.

Caudy, A. A., Ketting, R. F., Hammond, S. M., Denli, A. M., Bathoorn, A. M. P., Tops, B. B. J., Silva, J. M., Myers, M. M., Hannon, G. J., & Plasterk, R. H. A. (2003). A micrococcal nuclease homologue in RNAi effector complexes. *Nature*, 425, 411-414.

Chang, J. M., Di Tommaso, P., Notredame, C. (2014). TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol* 31(6), 1625-1637.

Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., & Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *PNAS*, 112(48), 14912–14917.

Charier, G., Couprie, J., Alpha-Bazin, B., Meyer, V., Quéméneur, E., Guérois, R., Callebaut, I., Gilquin, B., & Zinn-Justin, S. (2004). The tudor tandem of 53BP1: A new structural motif involved in DNA and RG-rich peptide binding. *Structure*, 12(9), 1551–1562.

Chen, C., Jin, J., James, D. A., Adams-Cioaba, M. A., Park, J. G., Guo, Y., Tenaglia, E., Xu, C., Gish, G., Min, J., & Pawson, T. (n.d.). Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. PNAS, 106(48), 20336–20341.

Chen, C., Nott, T. J., Jin, J., & Pawson, T. (2011). Deciphering arginine methylation: Tudor tells the tale. *Nature Reviews Molecular Cell Biology* 12(10), 629–642.

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Research*, 33, W72–W76.

Cheng, P., Huang, Y., Du, H., Li, C., Lv, Y., Ruan, R., Ye, H., Bian, C., You, X., Xu, J., Liang, X., Shi, Q., & Wei, Q. (2019). Draft genome and complete hox-cluster characterization of the sterlet sturgeon (*Acipenser ruthenus*). *Frontiers in Genetics*, 10, 776.

Cherif-Feildel, M., Kellner, K., Goux, D., Elie, N., Adeline, B., Lelong, C., & Heude Berthelin, C. (2019). Morphological and molecular criteria allow the identification of putative germ stem cells in a lophotrochozoan, the Pacific oyster *Crassostrea gigas*. *Histochemistry and Cell Biology*, 151(5), 419-433.

Chuma, S., Hosokawa, M., Kitamura, K., Kasai, S., Fujioka, M., Hiyoshi, M., Takamune, K., Noce, T., & Nakatsuji, N. (2006). *Tdrd1/Mtr-1*, a tudor-related gene, is essential for male germ-cell differentiation and nuagegerminal granule formation in mice. PNAS, 103(43), 15894–15899.

Cinalli, R. M., Rangan, P., & Lehmann, R. (2008). Germ Cells Are Forever. *Cell* 132(4), 559–562.

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18), 2938–2940.

Côté, J., & Richard, S. (2005). Tudor domains bind symmetrical dimethylated arginines. *Journal of Biological Chemistry*, 280(31), 28476–28483.

Cox, D. N., Chao, A., Baker, J., Chang, L., Qiao, D., & Lin, H. (1998). A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes & Development 12, 3715–3727.

Criscuolo, A., Gribaldo, S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10, 210

Cui, G., Park, S., Badeaux, A. I., Kim, D., Lee, J., Thompson, J. R., Yan, F., Kaneko, S., Yuan, Z., Botuyan, M. V., Bedford, M. T., Cheng, J. Q., & Mer, G. (2012). PHF20 is an effector protein of p53 double lysine methylation that stabilizes and activates p53. *Nature Structural and Molecular Biology*, 19(9), 916–924.

Czech, B., Munafò, M., Munafò, M., Ciabrelli, F., Eastwood, E. L., Fabry, M. H., Kneuss, E., & Hannon, G. J. (2018). piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annual Review of Genetics*, 52:131–57.

D'Orazio, F. M., Balwierz, P. J., González, A. J., Guo, Y., Hernández-Rodríguez, B., Wheatley, L., Jasiulewicz, A., Hadzhiev, Y., Vaquerizas, J. M., Cairns, B., Lenhard, B., & Müller, F. (2021). Germ cell differentiation requires Tdrd7-dependent chromatin and transcriptome reprogramming marked by germ plasm relocalization. *Developmental Cell*, 56(5), 641-656.e5.

Dailey, S. C., Planas, R. F., Espier, A. R., Garcia-Fernàndez, J., & Somorjai, I. M. L. (2016). Asymmetric distribution of *pl10* and *bruno2*, new members of a conserved core of early germline determinants in cephalochordates. *Frontiers in Ecology and Evolution*, 3, 156.

de Keuckelaere, E., Hulpiau, P., Saeys, Y., Berx, G., & van Roy, F. (2018). Nanos genes and their role in development and beyond. *Cellular and Molecular Life Sciences* 75(11), 1929–1946.

de Muyt, A., Pyatnitskaya, A., Andréani, J., Ranjha, L., Ramus, C., Laureau, R., Fernandez-Vega, A., Holoch, D., Girard, E., Govin, J., Margueron, R., Couté, Y., Cejka, P., Guérois, R., & Borde, V. (2018). A meiotic XPF–ERCC1-like complex recognizes joint molecule recombination intermediates to promote crossover formation. *Genes and Development*, 32(3–4), 283–296.

Deguchi, R., Takeda, N., & Stricker, S. A. (2011). Comparative biology of cAMP-induced germinal vesicle breakdown in marine invertebrate oocytes. *Molecular Reproduction and Development* 78(10–11), 708–725.

Del Val, E., Nasser, W., Abaibou, H., Reverchon, S. (2019). RecA and DNA recombination: a review of molecular mechanisms. *Biochem Soc* T 47, 1511-1531.

Devauchelle N (1990) Sviluppo Sessuale e Maturità di *Tapes philippinarum*. In: *Tapes philippinarum*: biologia e sperimentazione (First Ediction), Ente Sviluppo Agricolo Veneto (ESAV), 48–62

Dong, C., Nakagawa, R., Oyama, K., Yamamoto, Y., Zhang, W., Dong, A., Li, Y., Yoshimura, Y., Kamiya, H., Nakayama, J. I., Ueda, J., & Min, J. (2020). Structural basis for histone variant h3tk27me3 recognition by phf1 and phf19. *ELife*, 9, 1–21.

Du, K., Stöck, M., Kneitz, S., Klopp, C., Woltering, J. M., Adolfi, M. C., Feron, R., Prokopov, D., Makunin, A., Kichigin, I., Schmidt, C., Fischer, P., Kuhl, H., Wuertz, S., Gessner, J., Kloas, W., Cabau, C., Iampietro, C., Parrinello, H., … Schartl, M. (2020). The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nature Ecology and Evolution*, 4(6), 841–852.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195.

Ekman, D., Björklund, Å. K., & Elofsson, A. (2007). Quantification of the Elevated Rate of Domain Rearrangements in Metazoa. *Journal of Molecular Biology*, 372(5), 1337–1348.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238.

Eno, C., Hansen, C. L., & Pelegri, F. (2019). Aggregation, segregation, and dispersal of homotypic germ plasm RNPs in the early zebrafish embryo. *Developmental Dynamics*, 248(4), 306–318.

Ephrussi, A., Dickinson, L. K., & Lehmann, R. (1991). *oskar* Organizes the Germ Plasm and Directs Localization of the Posterior Determinant nanos. In *Cell* 66, 37-50.

Espinola-Lopez, J. M., & Tan, S. (2021). The Ada2/Ada3/Gcn5/Sgf29 histone acetyltransferase module. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1864, 194629 .

Ewen-Campen, B., Schwager, E. E., & Extavour, C. G. M. (2010). The molecular machinery of germ line specification. *Molecular Reproduction and Development* 77(1), 3–18.

Ewen-Campen, B., Srouji, J. R., Schwager, E. E., & Extavour, C. G. (2012). Oskar predates the evolution of germ plasm in insects. *Current Biology*, 22(23), 2278–2283.

Extavour, C. G. M. (2007). Evolution of the bilaterian germ line: Lineage origin and modulation of specification mechanisms. *Integrative and Comparative Biology*, 47(5), 770–785.

Extavour, C. G., & Akam, M. (2003). Mechanisms of germ cell specification across the metazoans: Epigenesis and preformation. *Development* 130(24), 5869–5884.

Falbo, L., Raspelli, E., Romeo, F., Fiorani, S., Pezzimenti, F., Casagrande, F., Costa, I., Parazzoli, D., & Costanzo, V. (2020). SSRP1-mediated histone H1 eviction promotes replication origin assembly and accelerated development. *Nature Communications*, 11, 1345.

Fei, Q., Shang, K., Zhang, J., Chuai, S., Kong, D., Zhou, T., Fu, S., Liang, Y., Li, C., Chen, Z., Zhao, Y., Yu, Z., Huang, Z., Hu, M., Ying, H., Chen, Z., Zhang, Y., Xing, F., Zhu, J., … Shou, J. (2015). Histone methyltransferase SETDB1 regulates liver cancer cell growth through methylation of P53. *Nature Communications*, 6, 8651.

Feist, S. W., Morris, D. J., Alama-Bermejo, G., & Holzer, A. S. (2015). Development and life cycles: Cellular processes in myxozoans. In *Myxozoan Evolution, Ecology and Development* (curated by Okamura, B., Gruhl, A., Bartholomew, J. L.). Springer International Publishing, 2015, 139–154.

Ferrández-Roldán, A., Martí-Solans, J., Cañestro, C., Albalat, R. (2019). *Oikopleura dioica*: An Emergent Chordate Model to Study the Impact of Gene Loss on the Evolution of the Mechanisms of Development. In *Evo-Devo: Nonmodel Species in Cell and Developmental* (curated by Tworzydlo, W., Bilinski, S.), Springer Nature Switzerland AG, 2019, Volume 68.

Fierro-Constaín, L., Schenkelaars, Q., Gazave, E., Haguenauer, A., Rocher, C., Ereskovsky, A., Borchiellini, C., & Renard, E. (2017). The conservation of the germline multipotency program, from sponges to vertebrates: A stepping stone to understanding the somatic and germline origins. *Genome Biology and Evolution*, 9(3), 474–488.

Forslund, S. K., Kaduk, M., Sonnhammer, E. L. L. (2019). Evolution of Protein Domain Architecture. In Evolutionary Genomics: Statistical and Computational Methods (Second Edition; curated by Anisimova M.), Humana Press, 2019.

Funayama, N. (2013). The stem cell system in demosponges: Suggested involvement of two types of cells: Archeocytes (active stem cells) and choanocytes (food-entrapping flagellated cells). *Development Genes and Evolution* 223(1–2), 23–38.

Funayama, N., Nakatsukasa, M., Mohri, K., Masuda, Y., & Agata, K. (2010). Piwi expression in archeocytes and choanocytes in demosponges: Insights into the stem cell system in demosponges. *Evolution and Development*, 12(3), 275–287.

Gao, X. D., Tachikawa, H., Sato, T., Jigami, Y., & Dean, N. (2005). Alg14 recruits Alg13 to the cytoplasmic face of the endoplasmic reticulum to form a novel bipartite UDP-N-acetylglucosamine transferase required for the second step of N-linked glycosylation. *Journal of Biological Chemistry*, 280(43), 36254–36262.

Gao, X., Ge, L., Shao, J., Su, C., Zhao, H., Saarikettu, J., Yao, X., Yao, Z., Silvennoinen, O., & Yang, J. (2010). Tudor-SN interacts with and co-localizes with G3BP in stress granules under stress conditions. *FEBS Letters*, 584(16), 3525–3532.

Gao, X., Zhao, X., Zhu, Y., He, J., Shao, J., Su, C., Zhang, Y., Zhang, W., Saarikettu, J., Silvennoinen, O., Yao, Z., & Yang, J. (2012). Tudor staphylococcal nuclease (Tudor-SN) participates in small ribonucleoprotein (snRNP) assembly via interacting with symmetrically dimethylated Sm proteins. *Journal of Biological Chemistry*, 287(22), 18130–18141.

Gazave, E., Béhague, J., Laplane, L., Guillou, A., Préau, L., Demilly, A., Balavoine, G., & Vervoort, M. (2013). Posterior elongation in the annelid *Platynereis dumerilii* involves stem cells molecularly related to primordial germ cells. *Developmental Biology*, 382(1), 246–267.

Gehrke, A. R., & Srivastava, M. (2016). Neoblasts and the evolution of whole-body regeneration. *Current Opinion in Genetics and Development* 40, 131–137.

Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L., Guiliano, D. B., Miranda-Saavedra, D., Angiuoli, S. v., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N. M., Wortman, J. R., Feldblyum, T., Tallon, L., Schatz, M., … Scott, A. L. (2007). Draft genome of the filarial nematode parasite *Brugia malayi*. *Science*, 317(5845), 1756–1760.

Gong, W., Liang, Q., Tong, Y., Perrett, S., & Feng, Y. (2021). Structural Insight into Chromatin Recognition by Multiple Domains of the Tumor Suppressor RBBP1. *Journal of Molecular Biology*, 433, 167224.

Gong, W., Wang, J., Perrett, S., & Feng, Y. (2014). Retinoblastoma-binding protein 1 has an interdigitated double tudor domain with DNA binding activity. *Journal of Biological Chemistry*, 289(8), 4882–4895.

Gonzalez, G. M., Hardwick, S. W., Maslen, S. L., Skehel, J. M., Holmqvist, E., Vogel, J., Bateman, A., Luisi, B. F., & William Broadhurst, R. (2017). Structure of the *Escherichia coli* ProQ RNA-binding protein. *RNA*, 23(5), 696-711.

Gosling E (2003) Bivalve molluscs: biology, ecology and culture. Fishing News Books, Blackwell Publishing, Oxford

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

Graham, W. J., Putnam, C. D., & Kolodner, R. D. (2018). DNA mismatch repair: Mechanisms and cancer genetics. In *Encyclopedia of Cancer* (curated by Boffetta, P., Hainaut, P.). Elsevier, 2018, 530–538.

Gribble, K. E., & Mark Welch, D. B. (2017). Genome-wide transcriptomics of aging in the rotifer *Brachionus manjavacas*, an emerging model system. *BMC Genomics*, 18(217).

Grohme, M. A., Schloissnig, S., Rozanski, A., Pippel, M., Young, G. R., Winkler, S., Brandl, H., Henry, I., Dahl, A., Powell, S., Hiller, M., Myers, E., & Rink, J. C. (2018). The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature*, 554(7690), 56–61.

Grudniewska, M., Mouton, S., Simanov, D., Beltman, F., Grelling, M., de Mulder, K., Arindrarto, W., Weissert, P. M., van der Elst, S., & Berezikov, E. (2016). Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife*, 5:e20607.

Gui, B., Han, X., Zhang, Y., Liang, J., Wang, D., Xuan, C., Yu, Z., & Shang, Y. (2012). Dimerization of ZIP promotes its transcriptional repressive function and biological activity. *International Journal of Biochemistry and Cell Biology*, 44(6), 886–895.

Gustafson, E. A., & Wessel, G. M. (2010). Vasa genes: Emerging roles in the germ line and in multipotent cells. *BioEssays* 32(7), 626–637.

Gutierrez-Beltran, E., Denisenko, T. v., Zhivotovsky, B., & Bozhkov, P. v. (2016). Tudor staphylococcal nuclease: Biochemistry and functions. *Cell Death and Differentiation* 23(11), 1739–1748.

Hahn, C., Fromm, B., & Bachmann, L. (2014). Comparative genomics of flatworms (Platyhelminthes) reveals shared genomic features of ecto- and endoparastic neodermata. *Genome Biology and Evolution*, 6(5), 1105–1117.

Handler, D., Olivieri, D., Novatchkova, M., Gruber, F. S., Meixner, K., Mechtler, K., Stark, A., Sachidanandam, R., & Brennecke, J. (2011). A systematic analysis of Drosophila TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *EMBO Journal*, 30(19), 3977–3993.

Hansen, C. L., & Pelegri, F. (2021). Primordial Germ Cell Specification in Vertebrate Embryos: Phylogenetic Distribution and Conserved Molecular Features of Preformation and Induction. *Frontiers in Cell and Developmental Biology*, 9, 730332.

Hashimoto, H., Hara, K., Hishiki, A., Kawaguchi, S., Shichijo, N., Nakamura, K., Unzai, S., Tamaru, Y., Shimizu, T., & Sato, M. (2010). Crystal structure of zinc-finger domain of Nanos and its functional implications. *EMBO Reports*, 11(11), 848–853.

Hashimoto, Y., Suzuki, H., Kageyama, Y., Yasuda, K., & Inoue, K. (2006). Bruno-like protein is localized to zebrafish germ plasm during the early cleavage stages. *Gene Expression Patterns*, 6(2), 201–205.

Hirose, T., Kawabuchi, M., Tamaru, T., Okumura, N., Nagai, K., & Okada, M. (2000). Identification of tudor repeat associator with PCTAIRE 2 (Trap) A novel protein that interacts with the N-terminal domain of PCTAIRE 2 in rat brain. *European. Journal of Biochemistry* 267, 2113-2121.

Hosokawa, M., Shoji, M., Kitamura, K., Tanaka, T., Noce, T., Chuma, S., & Nakatsuji, N. (2007). Tudor-related proteins TDRD1/MTR-1, TDRD6 and TDRD7/TRAP: Domain composition, intracellular localization, and function in male germ cells in mice. *Developmental Biology*, 301(1), 38–52.

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur, R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhou, Z., Huang, H., Chen, S., … Performed, Q. S. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *PNAS*, 115(24), 6249–6254.

Huo, L., Zhang, H., Huo, X., Yang, Y., Li, X., Yin, Y. (2017). pHMM-tree: phylogeny of profile hidden Markov models. *Bioinformatics* 33(7), 1093-1095.

Hur, J. H., van Doninck, K., Mandigo, M. L., & Meselson, M. (2009). Degenerate tetraploidy was established before bdelloid rotifer families diverged. *Molecular Biology and Evolution*, 26(2), 375–383.

Izumi, T., & Mellon, I. (2016). Base Excision Repair and Nucleotide Excision Repair. In *Genome Stability: From Virus to Human Application* (curated by Kovalchuk I., Kovalchuk, O.) Academic Press, 2016, pp. 275–302.

Jaron, K. S., Bast, J., Nowell, R. W., Ranallo-Benavidez, T. R., Robinson-Rechavi, M., & Schwander, T. (2021). Genomic Features of Parthenogenetic Animals. *The Journal of Heredity*, 112(1), 19–33.

Jeske, M., Bordi, M., Glatt, S., Müller, S., Rybin, V., Müller, C. W., & Ephrussi, A. (2015). The crystal structure of the *Drosophila* germline inducer Oskar identifies two domains with distinct Vasa Helicase- and RNA-binding activities. *Cell Reports*, 12(4), 587–598.

Jeske, M., Müller, C. W., & Ephrussi, A. (2017). The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes and Development*, 31(9), 939–952.

Jin, J., Xie, X., Chen, C., Gyoon Park, J., Stark, C., Andrew James, D., Olhovsky, M., Linding, R., Mao, Y., & Pawson, T. (n.d.). Eukaryotic Protein Domains as Functional Units of Cellular Evolution. *Science Signaling,* 2(98), ra76.

Johnson, A. D., & Alberio, R. (2015). Primordial germ cells: The first cell lineage or the last cells standing? *Development* 142(16), 2730–2739.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240.

Juliano, C. E., Swartz, S. Z., & Wessel, G. M. (2010). A conserved germline multipotency program. *Development* 137(24), 4113–4126.

Juliano, C. E., Voronina, E., Stack, C., Aldrich, M., Cameron, A. R., & Wessel, G. M. (2006). Germ line determinants are not localized early in sea urchin development, but do accumulate in the small micromere lineage. *Developmental Biology*, 300(1), 406–415.

Juliano, C., Wang, J., & Lin, H. (2011). Uniting germline and stem cells: The function of piwi proteins and the piRNA pathway in diverse organisms. *Annual Review of Genetics*, 45, 447–469.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780.

Kim, I. v., Riedelbauch, S., & Kuhn, C. D. (2020). The piRNA pathway in planarian flatworms: New model, new insights. *Biological Chemistry* 401(10), 1123–1141.

Kim, K. W. (2019). PIWI proteins and piRNAs in the nervous system. *Molecules and Cells* 42(12), 828–835.

Kim-Ha, J., Kerr, K., & Macdonald, P. M. (1995). Translational Regulation of oskar mRNA by Bruno, an Ovarian RNA-Binding Protein, Is Essential. In *Cell* 81, 403-412.

King, N., & Rokas, A. (2017). Embracing Uncertainty in Reconstructing Early Animal Evolution. *Current Biology* 27(19), R1081–R1088.

Kloc, M., Bilinski, S., & Etkin, L. D. (2004). The Balbiani Body and Germ Cell Determinants: 150 Years Later. *Current Topics in Developmental Biology*, 59.

Knoll, A. H. (2011). The multiple origins of complex multicellularity. *Annual Review of Earth and Planetary Sciences*, 39, 217–239.

Kobayashi, S., Yamada, M., Asaoka, M., Kitamura, T. (1996). Essential role of the posterior morphogen nanos for germline development in *Drosophila*. Nature 380, 708-711.

Kolb, S. J., Battle, D. J., & Dreyfuss, G. (2007). Molecular functions of the SMN complex. *Journal of Child Neurology* 22(8), 990–994.

Kori, S., Ferry, L., Matano, S., Jimenji, T., Kodera, N., Tsusaka, T., Matsumura, R., Oda, T., Sato, M., Dohmae, N., Ando, T., Shinkai, Y., Defossez, P. A., & Arita, K. (2019). Structure of the UHRF1 Tandem Tudor Domain Bound to a Methylated Non-histone Protein, LIG1, Reveals Rules for Binding and Regulation. *Structure*, *27*(3), 485-496.e7.

Kraaijeveld, K., Zwanenburg, B., Hubert, B., Vieira, C., de Pater, S., van Alphen, J. J. M., den Dunnen, J. T., & de Knijff, P. (2012). Transposon proliferation in an asexual parasitoid. *Molecular Ecology*, 21(16), 3898–3906.

Kranz, A. M., Tollenaere, A., Norris, B. J., Degnan, B. M., & Degnan, S. M. (2010). Identifying the germline in an equally cleaving mollusc: Vasa and Nanos expression during embryonic and larval development of the vetigastropod *Haliotis asinina*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 314 B(4), 267–279.

Krishna, S., Palakodeti, D., & Solana, J. (2019). Post-transcriptional regulation in planarian stem cells. In *Seminars in Cell and Developmental Biology* 87, 69–78.

Krishnakumar, P., Riemer, S., Perera, R., Lingner, T., Goloborodko, A., Khalifa, H., Bontems, F., Kaufholz, F., El-Brolosy, M. A., & Dosch, R. (2018). Functional equivalence of germ plasm organizers. *PLoS Genetics*, *14*(11) e1007696.

Ku, H. Y., & Lin, H. (2014). PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. *National Science Review* 1(2), 205–218.

Ku, H. Y., Gangaraju, V. K., Qi, H., Liu, N., & Lin, H. (2016). Tudor-SN Interacts with Piwi Antagonistically in Regulating Spermatogenesis but Synergistically in Silencing Transposons in *Drosophila*. *PLoS Genetics*, 12(1), e1005813.

Kumar, D. T., Susmita, B., Christy, J. P., Doss, C. G. P., Zayed, H. (2019). Elucidating the role of interacting residues of the MSH2-MSH6 complex in DNA repair mechanism: a computational approach. *Adv Protein Chem Str* 115, 325-350.

Labbé, R. M., Holowatyj, A., & Yang, Z.-Q. (2014). Histone lysine demethylase (KDM) subfamily 4: structures, functions and therapeutic potential. *American Journal Translationa Research* 6(1), 1-15.

Lachke, S. A., Alkuraya, F. S., Kneeland, S. C., Ohn, T., Aboukhalil, A., Howell, G. R., Saadi, I., Cavallesco, R., Yue, Y., Tsai, A. C.-H., Saidas Nair, K., Cosma, M. I., Smith, R. S., Hodges, E., Alfadhli, S. M., Al-Hajeri, A., Shamseldin, H. E., Behbehani, A., Hannon, G. J., … Maas, L. (2011). Mutations in the RNA Granule Component TDRD7 Cause Cataract and Glaucoma. *Science* 331, 1571-1576.

Lai, A. G., & Aboobaker, A. A. (2018). EvoRegen in animals: Time to uncover deep conservation or convergence of adult stem cell evolution and regenerative processes. *Developmental Biology* 433(2), 118–131.

Lancelot, N., Charier, G., Couprie, J., Duband-Goulet, I., Alpha-Bazin, B., Quémeneur, E., Ma, E., Marsolier-Kergoat, M. C., Ropars, V., Charbonnier, J. B., Miron, S., Craescu, C. T., Callebaut, I., Gilquin, B., & Zinn-justin, S. (2007). The checkpoint Saccharomyces cerevisiae Rad9 protein contains a tandem tudor domain that recognizes DNA. *Nucleic Acids Research*, 35(17), 5898–5912.

Lasko, P. (2013). The DEAD-box helicase Vasa: Evidence for a multiplicity of functions in RNA processes and developmental biology. In *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1829(8), 810–816.

Laumer, C. E., Fernández, R., Lemer, S., Combosch, D., Kocot, K. M., Riesgo, A., Andrade, S. C. S., Sterrer, W., Sørensen, M. v., & Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B: Biological Sciences*, 286, 20190831.

Lazzari, M., Bettini, S., & Franceschini, V. (2014). Immunocytochemical characterisation of olfactory ensheathing cells of zebrafish. *Journal of Anatomy*, 224(2), 192–206.

Leatherman, J. L., Levin, L., Boero, J., Jongens, T. A. (2002). *germ cell-less* Acts to Repress Transcription during the Establishment of the *Drosophila* Germ Cell Lineage. *Current Biology*, 12, 1681-1685.

Leighton, L. J., Wei, W., Marshall, P. R., Ratnu, V. S., Li, X., Zajaczkowski, E. L., Spadaro, P. A., Khandelwal, N., Kumar, A., & Bredy, T. W. (2019). Disrupting the hippocampal Piwi pathway enhances contextual fear memory in mice. *Neurobiology of Learning and Memory*, 161, 202–209.

Lesecque, Y., Mouchiroud, D., & Duret, L. (2013). GC-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution*, 30(6), 1409–1419.

Li, G. M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Research* 18(1), 85–98.

Li, L., & Xie, T. (2005). Stem cell niche: Structure and function. *Annual Review of Cell and Developmental Biology* 21, 605–631.

Li, R., Zhang, H., Yu, W., Chen, Y., Gui, B., Liang, J., Wang, Y., Sun, L., Yang, X., Zhang, Y., Shi, L., Li, Y., & Shang, Y. (2009). ZIP: A novel transcription repressor, represses EGFR oncogene and suppresses breast carcinogenesis. *EMBO Journal*, 28(18), 2763–2776.

Liang, C. C., Zhan, B., Yoshikawa, Y., Haas, W., Gygi, S. P., & Cohn, M. A. (2015). UHRF1 Is a sensor for DNA interstrand crosslinks and recruits FANCD2 to initiate the Fanconi Anemia pathway. *Cell Reports*, 10(12), 1947–1956.

Lim, R. S. M., & Kai, T. (2015). A piece of the pi(e): The diverse roles of animal piRNAs and their PIWI partners. *Seminars in Cell and Developmental Biology*, 47–48, 17–31.

Linder, B., Plöttner, O., Kroiss, M., Hartmann, E., Laggerbauer, B., Meister, G., Keidel, E., & Fischer, U. (2008). Tdrd3 is a novel stress granule-associated protein interacting with the Fragile-X syndrome protein FMRP. *Human Molecular Genetics*, 17(20), 3236–3246.

Liokatis, S., Edlich, C., Soupsana, K., Giannios, I., Panagiotidou, P., Tripsianes, K., Sattler, M., Georgatos, S. D., & Politou, A. S. (2012). Solution structure and molecular interactions of lamin B receptor Tudor domain. *Journal of Biological Chemistry*, 287(2), 1032–1042.

Liu, K., Chen, C., Guo, Y., Lam, R., Bian, C., Xu, C., Zhao, D. Y., Jin, J., Mackenzie, F., Pawson, T., & Min, J. (2010). Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. *PNAS*, 107(43), 18398–18403.

Liu, N., Han, H., & Lasko, P. (2009). Vasa promotes *Drosophila* germline stem cell differentiation by activating *mei-P26* translation by directly interacting with a (U)-rich motif in its 3' UTR. *Genes and Development*, 23(23), 2742–2752.

Liu, Q., Greimann, J. C., & Lima, C. D. (2006). Reconstitution, Activities, and Structure of the Eukaryotic RNA Exosome. *Cell*, 127(6), 1223–1237.

Livigni, A., Scorziello, A., Agnese, S., Adornetto, A., Carlucci, A., Garbi, C., Castaldo, I., Annunziato, L., Avvedimento, E. v, & Feliciello, A. (2006). Mitochondrial AKAP121 Links cAMP and src Signaling to Oxidative Metabolism. *Molecular Biology of the Cell*, 17, 263–271.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(550).

Lowry, O. H., Rosebrough, N. J., Farr, A. L., & Randall, R. J. (1951). Protein Measurement With The Folin Phenol Reagent. *The Journal of Biological Chemistry*, 193(1), 265-275.

Lu, R., & Wang, G. G. (2013). Tudor: A versatile family of histone methylation "readers." *Trends in Biochemical Sciences* 38(11), 546–555.

MacRae, S. L., McKnight Croken, M., Calder, R. B., Aliper, A., Milholland, B., White, R. R., Zhavoronkov, A., Gladyshev, V. N., Seluanov, A., Gorbunonva, V., Zhang, Z. D., Vijg, J. (2015). DNA repair in species with extreme lifespan differences. Aging, 7(12), 1171-1182.

Maurer-Stroh, S., Dickens, N. J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F., & Ponting, C. P. (1990). Tiggers and DNA transposon fossils in the human genome. *Update Trends in Biochemical Sciences* 215(2).

Meagher, M., Epling, L. B., & Enemark, E. J. (2019). DNA translocation mechanism of the MCM complex and implications for replication initiation. *Nature Communications*, 10:3117.

Mikhailov, K. v., Konstantinova, A. v., Nikitin, M. A., Troshin, P. v., Rusin, L. Y., Lyubetsky, V. A., Panchin, Y. v., Mylnikov, A. P., Moroz, L. L., Kumar, S., & Aleoshin, V. v. (2009). The origin of Metazoa: A transition from temporal to spatial cell differentiation. *BioEssays*, 31(7), 758–768.

Mikhailov, K. v., Slyusarev, G. S., Nikitin, M. A., Logacheva, M. D., Penin, A. A., Aleoshin, V. v., & Panchin, Y. v. (2016). The Genome of *Intoshia linei* Affirms Orthonectids as Highly Simplified Spiralians. *Current Biology*, 26(13), 1768–1774.

Milani, L., & Ghiselli, F. (2015). Mitochondrial activity in gametes and transmission of viable mtDNA. *Biology Direct*, 10(22).

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., Pecci, A., Maurizii, M. G., & Passamonti, M. (2015). The expression of a novel mitochondrially-encoded gene in gonadic precursors may drive paternal inheritance of mitochondria. *PLoS ONE*, 10(9): e0137468.

Milani, L., Pecci, A., Ghiselli, F., Passamonti, M., Bettini, S., Franceschini, V., & Maurizii, M. G. (2017). VASA expression suggests shared germ line dynamics in bivalve molluscs. *Histochemistry and Cell Biology*, 148(2), 157–171.

Milani, L., Pecci, A., Ghiselli, F., Passamonti, M., Lazzari, M., Franceschini, V., & Maurizii, M. G. (2018). Germ cell line during the seasonal sexual rest of clams: finding niches of cells for gonad renewal. *Histochemistry and Cell Biology*, 149(1), 105–110.

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, 8-15183.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5), 1530–1534.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419.

Mochizuki, K., Nishimiya-Fujisawa, C., & Fujisawa, T. (2001). Universal occurrence of the vasa-related genes among metazoans and their germline expression in *Hydra*. *Development Genes and Evolution*, 211(6), 299–308.

Mokrina, M., Nagasawa, K., Kanamori, M., Natsuike, M., & Osada, M. (2021). Seasonal composition of immature germ cells in the Yesso scallop identified by vasa-like gene (*my-vlg*) and protein expression, with evidence of irregular germ cell differentiation accompanied with a high mortality event. *Aquaculture Reports*, 19, 100613.

Mol, C. D., Harris, J. M., Mcintosh, E. M., & Tainer, J. A. (n.d.). Human dUTP pyrophosphatase: uracil recognition by a hairpin and active sites formed by three separate subunits. *Structure* 4(9).

Nagy, L. G. (2017). Evolution: Complex Multicellular Life with 5,500 Genes. *Current Biology* 27(12), R609–R612.

Nakahata, S., Kotani, T., Mita, K., Kawasaki, T., Katsu, Y., Nagahama, Y., & Yamashita, M. (2003). Involvement of *Xenopus* Pumilio in the translational regulation that is specific to cyclin B1 mRNA during oocyte maturation. *Mechanisms of Development*, 120(8), 865–880.

Nakamura, A., & Seydoux, G. (2008). Less is more: Specification of the germline by transcriptional repression. *Development* 135(23), 3817–3827.

Nakjang, S., Williams, T. A., Heinz, E., Watson, A. K., Foster, P. G., Sendra, K. M., Heaps, S. E., Hirt, R. P., & Embley, T. M. (2013). Reduction and expansion inmicrosporidian genome evolution: New insights from comparative genomics. *Genome Biology and Evolution*, 5(12), 2285–2303.

Nikolakaki, E., Mylonis, I., & Giannakouros, T. (2017). Lamin B receptor: Interplay between structure, function and localization. *Cells* 6(28).

Nowell, R. W., Almeida, P., Wilson, C. G., Smith, T. P., Fontaneto, D., Crisp, A., Micklem, G., Tunnacliffe, A., Boschetti, C., & Barraclough, T. G. (2018). Comparative genomics of bdelloid rotifers: Insights from desiccating and nondesiccating species. *PLoS Biology*, 16(4), e2004830.

Nowell, R. W., Wilson, C. G., Almeida, P., Schiffer, P. H., Fontaneto, D., Becks, L., Rodriguez, F., Arkhipova, I. R., & Barraclough, T. G. (2021). Evolutionary dynamics of transposable elements in bdelloid rotifers. *ELife*, 10, 1–86.

Obata, M., Sano, N., Kimata, S., Nagasawa, K., Yoshizaki, G., & Komaru, A. (2010). The proliferation and migration of immature germ cells in the mussel, *Mytilus galloprovincialis*: Observation of the expression pattern in the *M. galloprovincialis* vasa-like gene (*Myvlg*) by in situ hybridization. *Development Genes and Evolution*, 220(5–6), 139–149.

Önal, P., Grün, D., Adamidi, C., Rybak, A., Solana, J., Mastrobuoni, G., Wang, Y., Rahn, H. P., Chen, W., Kempa, S., Ziebold, U., & Rajewsky, N. (2012). Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. *EMBO Journal*, 31(12), 2755–2769.

Özpolat, B. D., & Bely, A. E. (2016). Developmental and molecular biology of annelid regeneration: a comparative review of recent studies. *Current Opinion in Genetics and Development* 40, 144–153.

Parisi, M., & Lin, H. (1999). The *Drosophila pumilio* Gene Encodes Two Functional Protein Isoforms That Play Multiple Roles in Germline Development, Gonadogenesis, Oogenesis and Embryogenesis. *Genetics* 153, 235–250.

Patil, V. S., Anand, A., Chakrabarti, A., & Kai, T. (2014). The Tudor domain protein Tapas, a homolog of the vertebrate Tdrd7, functions in the piRNA pathway to regulate retrotransposons in germline of *Drosophila melanogaster*. *BMC Biology*, 12(61).

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 417–419.

Pek, J. W., Anand, A., & Kai, T. (2012). Tudor domain proteins in development. *Development* 139(13), 2255–2266.

Perera, B. P. U., Tsai, Z. T. Y., Colwell, M. L., Jones, T. R., Goodrich, J. M., Wang, K., Sartor, M. A., Faulk, C., & Dolinoy, D. C. (2019). Somatic expression of piRNA and associated machinery in the mouse identifies short, tissue-specific piRNA. *Epigenetics*, 14(5), 504–521.

Ponting, C. P. (1997). Tudor domains in proteins that interact with RNA. Trends in Biochemical Sciences, 22(2), 51-52.

Poulin, R., & Randhawa, H. S. (2015). Evolution of parasitism along convergent lines: From ecology to genomics. *Parasitology* 142, S6–S15.

Rao, A. U., Carta, L. K., Lesuisse, E., & Hamza, I. (2005). Lack of heme synthesis in a free-living eukaryote. *PNAS*, 102(12), 4270–4275.

Rebscher, N., Zelada-González, F., Banisch, T. U., Raible, F., & Arendt, D. (2007). Vasa unveils a common origin of germ cells and of somatic stem cells from the posterior growth zone in the polychaete *Platynereis dumerilii*. *Developmental Biology*, 306(2), 599–611.

Reunov, A., Alexandrova, Y., Reunova, Y., Komkova, A., & Milani, L. (2019). Germ plasm provides clues on meiosis: The concerted action of germ plasm granules and mitochondria in gametogenesis of the clam *Ruditapes philippinarum*. *Zygote* 27(1), 5–16.

Richardson, R. T., Alekseev, O. M., Grossman, G., Widgren, E. E., Thresher, R., Wagner, E. J., Sullivan, K. D., Marzluff, W. F., & O'Rand, M. G. (2006). Nuclear autoantigenic sperm protein (NASP), a linker histone chaperone that is required for cell proliferation. *Journal of Biological Chemistry*, 281(30), 21526–21534.

Richter, D. J., & King, N. (2013). The genomic and cellular foundations of animal origins. *Annual Review of Genetics* 47, 509–537.

Rinkevich, Y., Rosner, A., Rabinowitz, C., Lapidot, Z., Moiseeva, E., & Rinkevich, B. (2010). Piwi positive cells that line the vasculature epithelium, underlie whole body regeneration in a basal chordate. *Developmental Biology*, 345(1), 94–104.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.

Robu, M. E., Inman, R. B., Cox M. M. (2001). RecA protein promotes the regression of stalled replication forks *in vitro*. *PNAS* 98(15), 8211-8218.

Rödelsperger, C., Streit, A., & Sommer, R. J. (2013). Structure, Function and Evolution of The Nematode Genome. In: *eLS*. John Wiley & Sons, Ltd: Chichester.

Rödelsperger, C. (2017). Comparative Genomics of Gene Loss and Gain in *Caenorhabditis* and Other Nematodes. In Comparative Genomics: Methods and Protocols (curated by Setubal J. C.), *Methods in Molecular Biology*, 1704, 419-432.

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acid Res* 47(W1), W5-W10.

Ruiz-Trillo, I., & Paps, J. (2016). Acoelomorpha: earliest branching bilaterians or deuterostomes? *Organisms Diversity and Evolution* 16(2), 391–399.

Ruiz-Trillo, I., Roger, A. J., Burger, G., Gray, M. W., & Lang, B. F. (2008). A phylogenomic investigation into the origin of Metazoa. *Molecular Biology and Evolution*, 25(4), 664–672.

Sabatella, M., Thijssen, K. L., Davó-Martínez, C., Vermeulen, W., & Lans, H. (2021). Tissue-Specific DNA Repair Activity of ERCC-1/XPF-1. *Cell Reports*, 34, 108608.

Sarkies, P., Selkirk, M. E., Jones, J. T., Blok, V., Boothby, T., Goldstein, B., Hanelt, B., Ardila-Garcia, A., Fast, N. M., Schiffer, P. M., Kraus, C., Taylor, M. J., Koutsovoulos, G., Blaxter, M. L., & Miska, E. A. (2015). Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. *PLoS Biology*, 13(2) e1002061.

Sasidharan, V., Lu, Y. C., Bansal, D., Dasari, P., Poduval, D., Seshasayee, A., Resch, A. M., Graveley, B. R., & Palakodeti, D. (2013). Identification of neoblast- And regeneration-specific miRNAs in the planarian *Schmidtea mediterranea*. *RNA*, 19(10), 1394–1404.

Sato, K., Iwasaki, Y. W., Shibuya, A., Carninci, P., Tsuchizawa, Y., Ishizu, H., Siomi, M. C., & Siomi, H. (2015). Krimper Enforces an Antisense Bias on piRNA Pools by Binding AGO3 in the *Drosophila* Germline. *Molecular Cell*, *59*(4), 553–563.

Schmitz, J. F., Zimmer, F., & Bornberg-Bauer, E. (2016). Mechanisms of transcription factor evolution in Metazoa. *Nucleic Acids Research*, 44(13), 6287–6297.

Schupbach, T., Wieschaus, E. (1986). Maternal-effect mutations altering the anterior-posterior pattern of the *Drosophila* embryo. *Roux's Archives of Developmental Biology* 195, 302-317.

Sebé-Pedrós, A., Degnan, B. M., & Ruiz-Trillo, I. (2017). The origin of Metazoa: A unicellular perspective. *Nature Reviews Genetics* 18(8), 498–512.

Selenko, P., Sprangers, R., Stier, G., Bühler, D., Fischer, U., & Sattler, M. (2001). SMN Tudor domain structure and its interaction with the Sm proteins. *Nature Structural Biology*, 8(1)-27-31.

Sengupta, M. S., & Boag, P. R. (2012). Germ granules and the control of mRNA translation. *IUBMB Life* 64(7), 586–594.

Sertic, S., Quadri, R., Lazzaro, F., Muzi-Falconi, M. (2020). EXO1: a tightly regulated nuclease. *DNA Repair* 93, 102929.

Seydoux, G., & Braun, R. E. (2006). Pathway to Totipotency: Lessons from Germ Cells. *Cell* 127(5), 891–904.

Shah, C., VanGompel, M. J. W., Naeem, V., Chen, Y., Lee, T., Angeloni, N., Wang, Y., & Xu, E. Y. (2010). Widespread presence of human BOULE homologs among animals and conservation of their ancient reproductive function. *PLoS Genetics*, 6(7), 1–16.

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst Biol* 51(3), 492-508.

Siomi, M. C., Mannen, T., & Siomi, H. (2010). How does the royal family of tudor rule the PIWI-interacting RNA pathway? *Genes and Development* 24(7), 636–646.

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: The vanguard of genome defence. *Nature Reviews Molecular Cell Biology* 12(4), 246–258.

Skinner, D. E., Rinaldi, G., Koziol, U., Brehm, K., & Brindley, P. J. (2014). How might flukes and tapeworms maintain genome integrity without a canonical piRNA pathway? *Trends in Parasitology* 30(3), 123–129.

Skorokhod, O. M., Gudkova, D. O., & Filonenko, V. V. (2011). Identification of novel TDRD7 isoforms. *Biopolymers and Cell* 27(6), 459–464.

Sogabe, S., Hatleberg, W. L., Kocot, K. M., Say, T. E., Stoupin, D., Roper, K. E., Fernandez-Valverde, S. L., Degnan, S. M., & Degnan, B. M. (2019). Pluripotency and the origin of animal multicellularity. *Nature*, 570(7762), 519–522.

Solana, J. (2013). Closing the circle of germline and stem cells: The Primordial Stem Cell hypothesis. *EvoDevo*, 4(2).

Srivastava, M., Begovic, E., Chapman, J., Putnam, N. H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M. L., Signorovitch, A. Y., Moreno, M. A., Kamm, K., Grimwood, J., Schmutz, J., Shapiro, H., Grigoriev, I. v., Buss, L. W., Schierwater, B., … Rokhsar, D. S. (2008). The *Trichoplax* genome and the nature of placozoans. *Nature*, 454(7207), 955–960.

Strasser, M. J., Mackenzie, N. C., Dumstrei, K., Nakkrasae, L. I., Stebler, J., & Raz, E. (2008). Control over the morphology and segregation of Zebrafish germ cell granules during embryonic development. *BMC Developmental Biology*, 8(58).

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, *6*(7).

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony. Version 4. Sinauer Associates, Sunderland, Massachusetts

Szitenberg, A., Cha, S., Opperman, C. H., Bird, D. M., Blaxter, M. L., & Lunt, D. H. (2016). Genetic drift, not life history or RNAi, determine long-Term evolution of transposable elements. *Genome Biology and Evolution* (9), 2964–2978.

Tadokoro, R., Sugio, M., Kutsuna, J., Tochinai, S., & Takahashi, Y. (2006). Early Segregation of Germ and Somatic Lineages during Gonadal Regeneration in the Annelid *Enchytraeus japonensis*. *Current Biology*, 16(10), 1012–1017.

Tanaka, T., Hosokawa, M., Vagin, V. v., Reuter, M., Hayashi, E., Mochizuki, A. L., Kitamura, K., Yamanaka, H., Kondoh, G., Okawa, K., Kuramochi-Miyagawa, S., Nakano, T., Sachidanandam, R., Hannon, G. J., Pillai, R. S., Nakatsuji, N., & Chuma, S. (2011). Tudor domain containing 7 (Tdrd7) is essential for dynamic ribonucleoprotein (RNP) remodeling of chromatoid bodies during spermatogenesis. *PNAS*, 108(26), 10579–10584.

Thangavel, S., Berti, M., Levikova, M., Pinto, C., Gomathinayagam, S., Vujanovic, M., Zellweger, R., Moore, H., Lee, E. H., Hendrickson, E. A., Cejka, P., Stewart, S., Lopes, M., & Vindigni, A.

(2015). DNA2 drives processing and restart of reversed replication forks in human cells. *Journal of Cell Biology*, 208(5), 545–562.

Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M. A., del Campo, J., Eme, L., Pérez-Cordón, G., Whipps, C. M., Nichols, K. M., Paley, R., Roger, A. J., Sitjà-Bobadilla, A., Donachie, S., & Ruiz-Trillo, I. (2015). Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology*, 25(18), 2404–2410.

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R. J., Fragoso, G., Sciutto, E., Aslett, M., Beasley, H., Bennett, H. M., Cai, J., Camicia, F., Clark, R., Cucher, M., de Silva, N., Day, T. A., … Valdes, V. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*, 496(7443), 57–63.

Tsai, L., & Barnea, G. (2014). A critical period defined by axon-targeting mechanisms in the murine olfactory bulb. *Science*, 344(6180), 197–200.

Tsutakawa, S. E., Sarker, A. H., Ng, C., Arvai, A. S., Shin, D. S., Shih, B., Jiang, S., Thwin, A. C., Tsai, M.-S., Willcox, A., Zong Her, M., Trego, K. S., Raetz, A. G., Rosenberg, D., Bacolla, A., Hammel, M., Griffith, J. D., Cooper, P. K., & Tainer, J. A. (2020). Human XPG nuclease structure, assembly, and activities with insights for neurodegeneration and cancer from pathogenic mutations. *PNAS*, 117(25), 14127–14138.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40(15), e115.

Vagin, V. v., Wohlschlegel, J., Qu, J., Jonsson, Z., Huang, X., Chuma, S., Girard, A., Sachidanandam, R., Hannon, G. J., & Aravin, A. A. (2009). Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes and Development*, 23(15), 1749–1762.

Viney, M. (2018). The genomic basis of nematode parasitism. *Briefings in Functional Genomics* 17(1), 8–14.

Voronina, E., Seydoux, G., Sassone-Corsi, P., & Nagamori, I. (2011). RNA granules in germ cells. *Cold Spring Harbor Perspectives in Biology* 3:a002774.

Wagner, D. E., Ho, J. J., & Reddien, P. W. (2012). Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. *Cell Stem Cell*, 10(3), 299–311.

Wagner, D. E., Wang, I. E., & Reddien, P. W. (2011). Clonogenic Neoblasts Are Pluripotent Adult Stem Cells That Underlie Planarian Regeneration. *Science* 332, 811-816

Wagner, G. P., & Lynch, V. J. (2008). The gene regulatory logic of transcription factor evolution. *Trends in Ecology and Evolution* 23(7), 377–385.

Weiner, K. X. B., Weiner, R. S., Maley, F., & Maleys, G. F. (1993). Primary Structure of Human Deoxycytidylate Deaminase and Overexpression of Its Functional Protein in *Escherichia coli. The Journal Of Biological Chemistry,* 268(17), 12983-12989.

Weinstein, S. B., & Kuris, A. M. (2016). Independent origins of parasitism in Animalia. *Biology Letters*, 12, 20160324 .

Weissman, A. (1892). *Das Keinplasma. Eine Theorie der Vererbung*.

Wessel, G. M. (2016). Germ Line Mechanics—And Unfinished Business. *Current Topics in Developmental Biology* 117, 553–566.

Whetstine, J. R., Nottke, A., Lan, F., Huarte, M., Smolikov, S., Chen, Z., Spooner, E., Li, E., Zhang, G., Colaiacovo, M., & Shi, Y. (2006). Reversal of Histone Lysine Trimethylation by the JMJD2 Family of Histone Demethylases. *Cell*, 125(3), 467–481.

Woodland, H. R. (2016). The Birth of Animal Development: Multicellularity and the Germline. In *Current Topics in Developmental Biology* 117, 609–630.

Wu, H., Zeng, H., Lam, R., Tempel, W., Amaya, M. F., Xu, C., Dombrovski, L., Qiu, W., Wang, Y., & Min, J. (2011). Structural and histone binding ability characterizations of human PWWP domains. *PLoS ONE*, 6(6), e18919.

Wudarski, J., Simanov, D., Ustyantsev, K., de Mulder, K., Grelling, M., Grudniewska, M., Beltman, F., Glazenburg, L., Demircan, T., Wunderer, J., Qi, W., Vizoso, D. B., Weissert, P. M., Olivieri, D., Mouton, S., Guryev, V., Aboobaker, A., Schärer, L., Ladurner, P., & Berezikov, E. (2017). Efficient transgenesis and annotated genome sequence of the regenerative flatworm model *Macrostomum lignano*. *Nature Communications*, 8, 2120.

Xie, T., Spradling, A.C (2000). A Niche Maintaining Germ Line Stem Cells in the *Drosophila* Ovary. Science 290, 328-330

Xue, X., Suvorov, A., Fujimoto, S., Dilman, A. R., Adams, B.J. (2021). Genome analysis of *Plectus murrayi*, a nematode from continental Antarctica. *G3* 11(1), jkaa045.

Yabuta, Y., Ohta, H., Abe, T., Kurimoto, K., Chuma, S., & Saitou, M. (2011). TDRD5 is required for retrotransposon silencing, chromatoid body assembly, and spermiogenesis in mice. *Journal of Cell Biology*, 192(5), 781–795.

Ying, M., & Chen, D. (2012). Tudor domain-containing proteins of *Drosophila melanogaster*. In *Development Growth and Differentiation* 54(1)-32-43.

Yuan, W., Al-Hadid, Q., Wang, Z., Shen, L., Cho, H., Wu, X., & Yang, Y. (2021). TDRD3 promotes DHX9 chromatin recruitment and R-loop resolution. *Nucleic Acids Research*, 49(15), 8573–8591.

Zadesenets, K. S., Schärer, L., & Rubtsov, N. B. (2017). New insights into the karyotype evolution of the free-living flatworm *Macrostomum lignano* (Platyhelminthes, Turbellaria). *Scientific Reports*, 7, 6066.

Zamparini, A. L., Davis, M. Y., Malone, C. D., Vieira, E., Zavadil, J., Sachidanandam, R., Hannon, G. J., & Lehmann, R. (2011). Vreteno, a gonad-specific protein, is essential for germline development and primary pirna biogenesis in *Drosophila*. *Development*, 138(18), 4039–4050.

Zarowiecki, M., & Berriman, M. (2015). What helminth genomes have taught us about parasite evolution. In *Parasitology* 142, S85–S97.

Zhu, X.-D., Niedernhofer, L., Kuster, B., Mann, M., Hoeijmakers, J. H., & de Lange, T. (2003). ERCC1/XPF Removes the 3 Overhang from Uncapped Telomeres and Represses Formation of Telomeric DNA-Containing Double Minute Chromosomes created at the end generated by leading strand DNA. *Molecular Cell* 12, 1489–1498.

Zouros, E., Ball, A. O., Saavedra, C., & Freeman$, K. R. (1994). An unusual type of mitochondrial DNA inheritance in the blue mussel Mytilus. In *Genetics* (Vol. 91).

Zverkov, O. A., Mikhailov, K. V., Isaev, S. V., Rusin, L. Y., Popova, O. V., Logacheva, M. D., Penin, A. A., Moroz, L. L., Panchin, Y. V., Lyubetsky, V. A., Aleoshin, V. V. (2019). Dicyemida and Orthonectida: two stories of body plan simplification. *Front Genet* 10(443).

Supplementary Material

The following section collects Supplementary Tables and Figures of all three chapters of the thesis. They are displayed in order, and the enumeration starts from 1 for every chapter, with a suffix corresponding to the letter of the chapter:

Chapter A: Supplementary Tables A1 to A12

Chapter B: Supplementary Tables B1 and B2

Chapter C: Supplementary Table C1 and Figures C1 to C3

Chapter A

Supplementary Table A1. Co-enriched germline-related Biological Processes GO terms shared by 5 or more species. GO terms were collapsed for semantic similarity (SimRel cut-off of 0.9; ReviGO server; see Materials and Methods). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group. N° of Species is the number of species that had that specific GO term enriched in germline-related samples at least twice as expected.

TermID	Name	Value
GO:0000003	**reproduction**	**6**
GO:0022414	**reproductive process**	**6**
GO:0044085	**cellular component biogenesis**	**9**
GO:2001251	**negative regulation of chromosome organization**	**5**
GO:0098813	**nuclear chromosome segregation**	**6**
GO:0006913	**nucleocytoplasmic transport**	**6**
GO:0006606	protein import into nucleus	5
GO:0051170	import into nucleus	5
GO:0007059	**chromosome segregation**	**7**
GO:0007017	**microtubule-based process**	**5**
GO:0006457	**protein folding**	**5**
GO:0007049	**cell cycle**	**5**
GO:0034660	**ncRNA metabolic process**	**9**
GO:0032259	**methylation**	**7**
GO:1901360	**organic cyclic compound metabolic process**	**8**
GO:0060249	**anatomical structure homeostasis**	**5**
GO:0051983	**regulation of chromosome segregation**	**5**
GO:0006260	**DNA replication**	**8**
GO:0034641	**cellular nitrogen compound metabolic process**	**9**
GO:0009057	**macromolecule catabolic process**	**6**
GO:0051726	**regulation of cell cycle**	**5**
GO:0031503	**protein-containing complex localization**	**5**
GO:1901990	**regulation of mitotic cell cycle phase transition**	**5**
GO:0000075	cell cycle checkpoint signaling	5
GO:0045930	negative regulation of mitotic cell cycle	5
GO:0007093	mitotic cell cycle checkpoint signaling	5

GO:0031570	DNA integrity checkpoint signaling	5
GO:0051052	**regulation of DNA metabolic process**	**5**
GO:0046483	**heterocycle metabolic process**	**7**
GO:0006725	**cellular aromatic compound metabolic process**	**8**
GO:0050658	**RNA transport**	**5**
GO:0051236	establishment of RNA localization	5
GO:0010467	**gene expression**	**7**
GO:0018193	**peptidyl-amino acid modification**	**7**
GO:0009262	**deoxyribonucleotide metabolic process**	**7**
GO:0006417	**regulation of translation**	**5**
GO:0009059	**macromolecule biosynthetic process**	**6**
GO:0019692	**deoxyribose phosphate metabolic process**	**6**
GO:0006261	**DNA-dependent DNA replication**	**8**
GO:0043631	**RNA polyadenylation**	**5**
GO:0034248	**regulation of cellular amide metabolic process**	**5**
GO:0010608	**posttranscriptional regulation of gene expression**	**5**
GO:0006383	**transcription by RNA polymerase III**	**5**
GO:0000280	**nuclear division**	**7**
GO:0000070	mitotic sister chromatid segregation	7
GO:0000819	sister chromatid segregation	6
GO:0140014	mitotic nuclear division	6
GO:0007062	sister chromatid cohesion	5
GO:0008213	**protein alkylation**	**6**
GO:0018205	**peptidyl-lysine modification**	**6**
GO:0006139	**nucleobase-containing compound metabolic process**	**7**
GO:0006281	**DNA repair**	**9**
GO:0006974	cellular response to DNA damage stimulus	8
GO:0043414	**macromolecule methylation**	**6**
GO:0006950	**response to stress**	**8**
GO:0009263	**deoxyribonucleotide biosynthetic process**	**7**
GO:0009221	pyrimidine deoxyribonucleotide biosynthetic process	6
GO:0046385	deoxyribose phosphate biosynthetic process	6
GO:0009394	2'-deoxyribonucleotide metabolic process	6
GO:0009219	pyrimidine deoxyribonucleotide metabolic process	6
GO:0009265	2'-deoxyribonucleotide biosynthetic process	6
GO:0071826	**ribonucleoprotein complex subunit organization**	**7**
GO:0048285	**organelle fission**	**6**
GO:0006403	**RNA localization**	**5**
GO:0001522	**pseudouridine synthesis**	**6**
GO:0000723	**telomere maintenance**	**5**
GO:0070647	**protein modification by small protein conjugation or removal**	**5**
GO:0006259	**DNA metabolic process**	**8**
GO:0034504	**protein localization to nucleus**	**5**
GO:0009130	**pyrimidine nucleoside monophosphate biosynthetic process**	**6**
GO:0032200	**telomere organization**	**5**
GO:0016071	**mRNA metabolic process**	**7**
GO:0034645	**cellular macromolecule biosynthetic process**	**6**
GO:0051276	**chromosome organization**	**7**
GO:0000226	**microtubule cytoskeleton organization**	**5**
GO:0016070	**RNA metabolic process**	**7**
GO:0022613	**ribonucleoprotein complex biogenesis**	**8**
GO:0009451	**RNA modification**	**6**
GO:0006221	**pyrimidine nucleotide biosynthetic process**	**6**
GO:0000413	**protein peptidyl-prolyl isomerization**	**5**
GO:0018208	**peptidyl-proline modification**	**5**
GO:0090304	**nucleic acid metabolic process**	**6**
GO:0065004	**protein-DNA complex assembly**	**5**
GO:0034470	**ncRNA processing**	**8**
GO:0051169	**nuclear transport**	**6**
GO:0071824	**protein-DNA complex subunit organization**	**5**

GO:0000724	**double-strand break repair via homologous recombination**	**6**
GO:0006479	**protein methylation**	**6**
GO:0044265	**cellular macromolecule catabolic process**	**6**
GO:0006270	**DNA replication initiation**	**8**
GO:0045005	**DNA-dependent DNA replication maintenance of fidelity**	**5**
GO:0000725	**recombinational repair**	**6**
GO:0051306	**mitotic sister chromatid separation**	**5**
GO:0006996	**organelle organization**	**7**
GO:0051783	**regulation of nuclear division**	**5**
GO:0006298	**mismatch repair**	**6**
GO:0042254	**ribosome biogenesis**	**6**
GO:0006364	rRNA processing	6
GO:0008380	**RNA splicing**	**6**
GO:0006396	**RNA processing**	**8**
GO:0006310	**DNA recombination**	**5**
GO:0071103	**DNA conformation change**	**7**
GO:0030488	**tRNA methylation**	**5**
GO:0006302	**double-strand break repair**	**7**
GO:0006323	**DNA packaging**	**6**
GO:0022402	**cell cycle process**	**5**
GO:0006289	**nucleotide-excision repair**	**6**
GO:0006397	**mRNA processing**	**7**
GO:0007091	**metaphase/anaphase transition of mitotic cell cycle**	**5**
GO:0044784	**metaphase/anaphase transition of cell cycle**	**5**
GO:0009129	**pyrimidine nucleoside monophosphate metabolic process**	**6**
GO:0000278	**mitotic cell cycle**	**5**
GO:0016072	**rRNA metabolic process**	**7**
GO:0034622	**cellular protein-containing complex assembly**	**7**
GO:0065003	protein-containing complex assembly	7
GO:0033554	**cellular response to stress**	**9**
GO:0008033	**tRNA processing**	**8**
GO:0007281	**germ cell development**	**5**
GO:0045786	**negative regulation of cell cycle**	**5**
GO:1903047	**mitotic cell cycle process**	**5**
GO:0006400	**tRNA modification**	**6**
GO:0006325	**chromatin organization**	**6**
GO:0007346	**regulation of mitotic cell cycle**	**5**
GO:0006402	**mRNA catabolic process**	**5**
GO:0001510	**RNA methylation**	**5**
GO:0022618	**ribonucleoprotein complex assembly**	**7**
GO:0018216	**peptidyl-arginine methylation**	**5**
GO:0044772	**mitotic cell cycle phase transition**	**5**
GO:0030163	**protein catabolic process**	**5**
GO:0006399	**tRNA metabolic process**	**7**
GO:0000387	**spliceosomal snRNP assembly**	**6**
GO:0000398	**mRNA splicing, via spliceosome**	**6**
GO:0000375	RNA splicing, via transesterification reactions	6
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	6
GO:0033047	**regulation of mitotic sister chromatid segregation**	**5**
GO:0030071	regulation of mitotic metaphase/anaphase transition	5
GO:1905818	regulation of chromosome separation	5
GO:0033045	regulation of sister chromatid segregation	5
GO:1902099	regulation of metaphase/anaphase transition of cell cycle	5
GO:0010965	regulation of mitotic sister chromatid separation	5
GO:0007088	**regulation of mitotic nuclear division**	**5**
GO:0003006	**developmental process involved in reproduction**	**5**
GO:0006220	pyrimidine nucleotide metabolic process	6
GO:0072528	pyrimidine-containing compound biosynthetic process	5
GO:0050657	**nucleic acid transport**	**5**

Supplementary Table A2. Co-enriched germline-related Molecular Functions GO terms shared by 5 or more species. GO terms were collapsed for semantic similarity (SimRel cut-off of 0.9; ReviGO server; see Materials and Methods). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group. N° of Species is the number of species that had that specific GO term enriched in germline-related samples at least twice as expected.

TermID	Name	Value
GO:0003712	**transcription coregulator activity**	**5**
GO:0004518	**nuclease activity**	**9**
GO:0030983	**mismatched DNA binding**	**7**
GO:0140104	**molecular carrier activity**	**5**
GO:0017056	**structural constituent of nuclear pore**	**5**
GO:0008641	**ubiquitin-like modifier activating enzyme activity**	**5**
GO:0016866	**intramolecular transferase activity**	**5**
GO:0031267	**small GTPase binding**	**6**
GO:0016779	**nucleotidyltransferase activity**	**8**
GO:0140098	**catalytic activity, acting on RNA**	**8**
GO:0016853	**isomerase activity**	**5**
GO:0016874	**ligase activity**	**5**
GO:0005524	**ATP binding**	**6**
GO:0003729	**mRNA binding**	**5**
GO:0034212	**peptide N-acetyltransferase activity**	**5**
GO:0003684	**damaged DNA binding**	**6**
GO:0003723	**RNA binding**	**8**
GO:0008168	**methyltransferase activity**	**8**
GO:0003697	**single-stranded DNA binding**	**5**
GO:0016741	**transferase activity, transferring one-carbon groups**	**8**
GO:0003690	**double-stranded DNA binding**	**5**
GO:0008017	**microtubule binding**	**6**
GO:0046982	**protein heterodimerization activity**	**5**
GO:0003677	**DNA binding**	**7**
GO:0016273	**arginine N-methyltransferase activity**	**5**
GO:0004402	**histone acetyltransferase activity**	**5**
GO:0061733	peptide-lysine-N-acetyltransferase activity	5
GO:0008094	**ATP-dependent activity, acting on DNA**	**7**
GO:0016888	**endodeoxyribonuclease activity, producing 5'-phosphomonoesters**	**6**
GO:0009982	**pseudouridine synthase activity**	**5**
GO:0004386	**helicase activity**	**6**
GO:0003755	**peptidyl-prolyl cis-trans isomerase activity**	**5**
GO:0016859	**cis-trans isomerase activity**	**5**
GO:0140101	**catalytic activity, acting on a tRNA**	**6**
GO:0034061	**DNA polymerase activity**	**5**
GO:0003678	**DNA helicase activity**	**7**
GO:0008276	**protein methyltransferase activity**	**6**
GO:0140097	**catalytic activity, acting on DNA**	**7**
GO:0043139	**5'-3' DNA helicase activity**	**5**
GO:0004536	**deoxyribonuclease activity**	**5**
GO:0008170	**N-methyltransferase activity**	**6**
GO:0008173	**RNA methyltransferase activity**	**6**
GO:0030554	**adenyl nucleotide binding**	**6**
GO:0004527	**exonuclease activity**	**7**
GO:0016274	**protein-arginine N-methyltransferase activity**	**5**
GO:0015631	**tubulin binding**	**6**
GO:0008757	**S-adenosylmethionine-dependent methyltransferase activity**	**6**
GO:0008175	**tRNA methyltransferase activity**	**5**
GO:0004519	**endonuclease activity**	**9**
GO:0032559	**adenyl ribonucleotide binding**	**6**

Supplementary Table A3. *Brachionus manjavacas* germline-related enriched GO terms. GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0009611	**response to wounding**	**GO:0032879**	**regulation of localization**
GO:0023052	**signaling**	**GO:0010646**	**regulation of cell communication**
GO:0030030	**cell projection organization**	**GO:0023051**	**regulation of signaling**
GO:0030431	**sleep**	**GO:0031503**	**protein-containing complex localization**
GO:0032501	**multicellular organismal process**	**GO:0006836**	**neurotransmitter transport**
GO:0032502	**developmental process**	**GO:0030705**	**cytoskeleton-dependent intracellular transport**
GO:0040007	**growth**	**GO:0035694**	**mitochondrial protein catabolic process**
GO:0040011	**locomotion**	**GO:0051674**	**localization of cell**
GO:0048856	**anatomical structure development**	**GO:0044091**	**membrane biogenesis**
GO:0022008	neurogenesis	**GO:0042060**	**wound healing**
GO:0009888	tissue development	**GO:0097479**	**synaptic vesicle localization**
GO:0007399	nervous system development	GO:0097480	establishment of synaptic vesicle localization
GO:0030154	cell differentiation	GO:0048489	synaptic vesicle transport
GO:0048869	cellular developmental process	**GO:0007186**	**G protein-coupled receptor signaling pathway**
GO:0048666	neuron development	**GO:0120036**	**plasma membrane bounded cell projection organization**
GO:0048731	system development	GO:0044782	cilium organization
GO:0007275	multicellular organism development	GO:0060271	cilium assembly
GO:0030182	neuron differentiation	GO:0042073	intraciliary transport
GO:0048699	generation of neurons	GO:0030031	cell projection assembly
GO:0050804	**modulation of chemical synaptic transmission**	GO:0120031	plasma membrane bounded cell projection assembly
GO:0032222	regulation of synaptic transmission, cholinergic	GO:0035082	axoneme assembly
GO:0007018	**microtubule-based movement**	GO:0010970	transport along microtubule
GO:0060294	cilium movement involved in cell motility	**GO:0009100**	**glycoprotein metabolic process**
GO:0001578	microtubule bundle formation	**GO:0007267**	**cell-cell signaling**
GO:0099111	microtubule-based transport	**GO:0070925**	**organelle assembly**
GO:0060285	cilium-dependent cell motility	**GO:0099504**	**synaptic vesicle cycle**
GO:0003341	cilium movement	GO:0016079	synaptic vesicle exocytosis
GO:0001539	cilium or flagellum-dependent cell motility	**GO:0099003**	**vesicle-mediated transport in synapse**
GO:0000226	microtubule cytoskeleton organization	**GO:0045055**	**regulated exocytosis**
GO:0048870	cell motility	GO:0017156	calcium-ion regulated exocytosis
GO:0034220	**ion transmembrane transport**	GO:1990504	dense core granule exocytosis
GO:0098662	inorganic cation transmembrane transport	**GO:0071709**	**membrane assembly**
GO:0030001	metal ion transport	GO:0007009	plasma membrane organization
GO:0070085	**glycosylation**	GO:0070836	caveola assembly
GO:0032409	**regulation of transporter activity**	GO:0044857	plasma membrane raft organization
GO:0051049	**regulation of transport**	GO:0001765	membrane raft assembly
GO:0043266	regulation of potassium ion transport	GO:0044854	plasma membrane raft assembly
GO:0043270	positive regulation of ion transport	**GO:0008277**	**regulation of G protein-coupled receptor signaling pathway**
GO:1904064	positive regulation of cation transmembrane transport	**GO:0031099**	**regeneration**
GO:0051050	positive regulation of transport	**GO:0099177**	**regulation of trans-synaptic signaling**
GO:0034765	regulation of ion transmembrane transport	**GO:0031579**	**membrane raft organization**
GO:0034764	positive regulation of transmembrane transport	**GO:0006813**	**potassium ion transport**
GO:0034762	regulation of transmembrane transport	GO:0071805	potassium ion transmembrane transport
GO:0032412	regulation of ion transmembrane transporter activity	**GO:0048589**	**developmental growth**
GO:0043269	regulation of ion transport	GO:0042246	tissue regeneration
GO:1904062	regulation of cation transmembrane transport	**GO:0006816**	**calcium ion transport**
GO:0010959	regulation of metal ion transport	**GO:0007268**	**chemical synaptic transmission**
GO:1901016	regulation of potassium ion transmembrane transporter activity	GO:0023061	signal release
GO:1903818	positive regulation of voltage-gated potassium channel activity	GO:0007271	synaptic transmission, cholinergic
GO:0032411	positive regulation of transporter activity	GO:0099536	synaptic signaling

GO:0043268	positive regulation of potassium ion transport	GO:0007269	neurotransmitter secretion
GO:2001259	positive regulation of cation channel activity	GO:0099643	signal release from synapse
GO:0034767	positive regulation of ion transmembrane transport	GO:0099537	trans-synaptic signaling
GO:1901018	positive regulation of potassium ion transmembrane transporter activity	GO:0098916	anterograde trans-synaptic signaling
GO:2001257	regulation of cation channel activity	**GO:0006486**	**protein glycosylation**
GO:0032414	positive regulation of ion transmembrane transporter activity	GO:0043413	macromolecule glycosylation
GO:1901379	regulation of potassium ion transmembrane transport	GO:0009101	glycoprotein biosynthetic process
GO:1901381	positive regulation of potassium ion transmembrane transport	**GO:0006928**	**movement of cell or subcellular component**
GO:0022898	regulation of transmembrane transporter activity	**GO:0007017**	**microtubule-based process**

Supplementary Table A4. ***Caenorhabditis elegans*** **germline-related enriched GO terms.** GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0006281	**DNA repair**	**GO:0006418**	**tRNA aminoacylation for protein translation**
GO:0006298	mismatch repair	GO:0043039	tRNA aminoacylation
GO:0006289	nucleotide-excision repair	**GO:0043038**	**amino acid activation**
GO:0033554	cellular response to stress	**GO:0009211**	**pyrimidine deoxyribonucleoside triphosphate metabolic process**
GO:0048569	**post-embryonic animal organ development**	GO:0009200	deoxyribonucleoside triphosphate metabolic process
GO:2000026	**regulation of multicellular organismal development**	GO:0009147	pyrimidine nucleoside triphosphate metabolic process
GO:0048580	regulation of post-embryonic development	GO:0009149	pyrimidine nucleoside triphosphate catabolic process
GO:0051241	negative regulation of multicellular organismal process	GO:0009204	deoxyribonucleoside triphosphate catabolic process
GO:0016999	**antibiotic metabolic process**	**GO:0061062**	**regulation of nematode larval development**
GO:0006083	**acetate metabolic process**	GO:0040027	negative regulation of vulval development
GO:0009057	**macromolecule catabolic process**	GO:0048581	negative regulation of post-embryonic development
GO:0042592	**homeostatic process**	GO:0061064	negative regulation of nematode larval development
GO:0051239	**regulation of multicellular organismal process**	**GO:0009221**	**pyrimidine deoxyribonucleotide biosynthetic process**
GO:0050793	**regulation of developmental process**	GO:0006221	pyrimidine nucleotide biosynthetic process
GO:0006417	**regulation of translation**	GO:0006220	pyrimidine nucleotide metabolic process
GO:0016180	**snRNA processing**	GO:0046081	dUTP catabolic process
GO:0009165	**nucleotide biosynthetic process**	GO:0046386	deoxyribose phosphate catabolic process
GO:0046390	ribose phosphate biosynthetic process	GO:0009263	deoxyribonucleotide biosynthetic process
GO:0009123	nucleoside monophosphate metabolic process	GO:0046078	dUMP metabolic process
GO:0009124	nucleoside monophosphate biosynthetic process	GO:0006226	dUMP biosynthetic process
GO:1901293	nucleoside phosphate biosynthetic process	GO:0046080	dUTP metabolic process
GO:0009152	purine ribonucleotide biosynthetic process	GO:0046385	deoxyribose phosphate biosynthetic process
GO:0009260	ribonucleotide biosynthetic process	GO:0009394	2'-deoxyribonucleotide metabolic process
GO:0006414	**translational elongation**	GO:0009219	pyrimidine deoxyribonucleotide metabolic process
GO:0016073	**snRNA metabolic process**	GO:0009223	pyrimidine deoxyribonucleotide catabolic process
GO:0034248	**regulation of cellular amide metabolic process**	GO:0009213	pyrimidine deoxyribonucleoside triphosphate catabolic process
GO:0032787	**monocarboxylic acid metabolic process**	GO:0009265	2'-deoxyribonucleotide biosynthetic process
GO:0010608	**posttranscriptional regulation of gene expression**	**GO:0044265**	**cellular macromolecule catabolic process**
GO:0002164	**larval development**	GO:0051603	proteolysis involved in cellular protein catabolic process
GO:0006950	**response to stress**	GO:0030163	protein catabolic process
GO:0006259	**DNA metabolic process**	GO:0044257	cellular protein catabolic process
GO:0040028	**regulation of vulval development**	**GO:0002119**	**nematode larval development**
GO:0019692	**deoxyribose phosphate metabolic process**	GO:0040025	vulval development
GO:0009791	**post-embryonic development**	**GO:0006244**	**pyrimidine nucleotide catabolic process**
GO:0048513	animal organ development	GO:0034655	nucleobase-containing compound catabolic process
GO:0051093	**negative regulation of developmental process**		

Supplementary Table A5. ***Clytia hemisphaerica*** **germline-related enriched GO terms.** GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0000003	**reproduction**	**GO:0035825**	**homologous recombination**
GO:0006260	**DNA replication**	**GO:0046939**	**nucleotide phosphorylation**
GO:0030162	**regulation of proteolysis**	**GO:0016485**	**protein processing**
GO:0051321	**meiotic cell cycle**	GO:0006465	signal peptide processing
GO:1903046	meiotic cell cycle process	**GO:0044283**	**small molecule biosynthetic process**
GO:0022414	reproductive process	**GO:0051604**	**protein maturation**
GO:0007131	reciprocal meiotic recombination	**GO:0006261**	**DNA-dependent DNA replication**
GO:0061982	meiosis I cell cycle process	**GO:0019692**	**deoxyribose phosphate metabolic process**
GO:0072657	**protein localization to membrane**	**GO:0046434**	**organophosphate catabolic process**
GO:0006457	**protein folding**	**GO:0007005**	**mitochondrion organization**
GO:0034622	**cellular protein-containing complex assembly**	**GO:0034404**	**nucleobase-containing small molecule biosynthetic process**
GO:0043248	proteasome assembly	**GO:0000280**	**nuclear division**
GO:0033108	mitochondrial respiratory chain complex assembly	**GO:1901292**	**nucleoside phosphate catabolic process**
GO:0017004	cytochrome complex assembly	GO:0009166	nucleotide catabolic process
GO:0071826	ribonucleoprotein complex subunit organization	**GO:0048285**	**organelle fission**
GO:0022618	ribonucleoprotein complex assembly	**GO:0016052**	**carbohydrate catabolic process**
GO:0000387	spliceosomal snRNP assembly	**GO:0009132**	**nucleoside diphosphate metabolic process**
GO:0065003	protein-containing complex assembly	**GO:0046394**	**carboxylic acid biosynthetic process**
GO:0019674	**NAD metabolic process**	GO:0072330	monocarboxylic acid biosynthetic process
GO:0008380	**RNA splicing**	GO:0016053	organic acid biosynthetic process
GO:0042866	**pyruvate biosynthetic process**	**GO:0000398**	**mRNA splicing, via spliceosome**
GO:0006760	**folic acid-containing compound metabolic process**	GO:0006397	mRNA processing
GO:0009396	folic acid-containing compound biosynthetic process	GO:0000375	RNA splicing, via transesterification reactions
GO:0042559	pteridine-containing compound biosynthetic process	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0046653	tetrahydrofolate metabolic process	**GO:0006096**	**glycolytic process**
GO:0072524	**pyridine-containing compound metabolic process**	GO:0009185	ribonucleoside diphosphate metabolic process
GO:0042558	**pteridine-containing compound metabolic process**	GO:0009135	purine nucleoside diphosphate metabolic process
GO:0072527	**pyrimidine-containing compound metabolic process**	GO:0046031	ADP metabolic process
GO:0019359	**nicotinamide nucleotide biosynthetic process**	GO:0006165	nucleoside diphosphate phosphorylation
GO:0009165	nucleotide biosynthetic process	GO:0009179	purine ribonucleoside diphosphate metabolic process
GO:0009130	pyrimidine nucleoside monophosphate biosynthetic process	GO:0006757	ATP generation from ADP
GO:0009129	pyrimidine nucleoside monophosphate metabolic process	**GO:0008535**	**respiratory chain complex IV assembly**
GO:0006221	pyrimidine nucleotide biosynthetic process	GO:0033617	mitochondrial cytochrome c oxidase assembly
GO:0009221	pyrimidine deoxyribonucleotide biosynthetic process	**GO:0006090**	**pyruvate metabolic process**
GO:1901293	nucleoside phosphate biosynthetic process	**GO:0006270**	**DNA replication initiation**
GO:0006220	pyrimidine nucleotide metabolic process	**GO:0016226**	**iron-sulfur cluster assembly**
GO:0019362	pyridine nucleotide metabolic process	**GO:0031163**	**metallo-sulfur cluster assembly**
GO:0072528	pyrimidine-containing compound biosynthetic process		
GO:0046385	deoxyribose phosphate biosynthetic process		
GO:0072525	pyridine-containing compound biosynthetic process		
GO:0019363	pyridine nucleotide biosynthetic process		
GO:0009394	2'-deoxyribonucleotide metabolic process		
GO:0009219	pyrimidine deoxyribonucleotide metabolic process		
GO:0009435	NAD biosynthetic process		
GO:0046496	nicotinamide nucleotide metabolic process		
GO:0009265	2'-deoxyribonucleotide biosynthetic process		

Supplementary Table A6. *Danio rerio* germline-related enriched GO terms. GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0000003	**reproduction**	**GO:0015931**	**nucleobase-containing compound transport**
GO:0022402	**cell cycle process**	**GO:0006281**	**DNA repair**
GO:1903047	mitotic cell cycle process	GO:0000725	recombinational repair
GO:0000278	mitotic cell cycle	GO:0006310	DNA recombination
GO:0000819	sister chromatid segregation	GO:0006302	double-strand break repair
GO:0000910	cytokinesis	GO:0033554	cellular response to stress
GO:0000281	mitotic cytokinesis	GO:0006974	cellular response to DNA damage stimulus
GO:0044772	mitotic cell cycle phase transition	**GO:0019220**	**regulation of phosphate metabolic process**
GO:0007076	mitotic chromosome condensation	**GO:0000290**	**deadenylation-dependent decapping of nuclear-transcribed mRNA**
GO:1902850	microtubule cytoskeleton organization involved in mitosis	GO:0000288	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
GO:0008608	attachment of spindle microtubules to kinetochore	**GO:0009130**	**pyrimidine nucleoside monophosphate biosynthetic process**
GO:0007091	metaphase/anaphase transition of mitotic cell cycle	GO:0009129	pyrimidine nucleoside monophosphate metabolic process
GO:0044784	metaphase/anaphase transition of cell cycle	GO:0009177	pyrimidine deoxyribonucleoside monophosphate biosynthetic process
GO:0098813	nuclear chromosome segregation	GO:0009176	pyrimidine deoxyribonucleoside monophosphate metabolic process
GO:0000086	G2/M transition of mitotic cell cycle	GO:0009157	deoxyribonucleoside monophosphate biosynthetic process
GO:0044839	cell cycle G2/M phase transition	**GO:0032784**	**regulation of DNA-templated transcription, elongation**
GO:0140014	mitotic nuclear division	**GO:0051174**	**regulation of phosphorus metabolic process**
GO:0000070	mitotic sister chromatid segregation	**GO:0043414**	**macromolecule methylation**
GO:0000280	nuclear division	**GO:0006261**	**DNA-dependent DNA replication**
GO:0032504	**multicellular organism reproduction**	**GO:0060249**	**anatomical structure homeostasis**
GO:0022412	cellular process involved in reproduction in multicellular organism	**GO:0042761**	**very long-chain fatty acid biosynthetic process**
GO:0048477	oogenesis	**GO:0006396**	**RNA processing**
GO:0022414	reproductive process	**GO:0050658**	**RNA transport**
GO:0019953	sexual reproduction	GO:0051030	snRNA transport
GO:0044703	multi-organism reproductive process	GO:0050657	nucleic acid transport
GO:0007292	female gamete generation	GO:0051236	establishment of RNA localization
GO:0003006	developmental process involved in reproduction	**GO:0051383**	**kinetochore organization**
GO:0007281	germ cell development	**GO:0000038**	**very long-chain fatty acid metabolic process**
GO:0007283	spermatogenesis	**GO:0007254**	**JNK cascade**
GO:0048232	male gamete generation	**GO:0034508**	**centromere complex assembly**
GO:0007276	gamete generation	GO:0051382	kinetochore assembly
GO:0048609	multicellular organismal reproductive process	GO:0065004	protein-DNA complex assembly
GO:0051726	**regulation of cell cycle**	**GO:0006403**	**RNA localization**
GO:0008037	**cell recognition**	**GO:0051306**	**mitotic sister chromatid separation**
GO:0006913	**nucleocytoplasmic transport**	**GO:0001522**	**pseudouridine synthesis**
GO:0007059	**chromosome segregation**	**GO:0034502**	**protein localization to chromosome**
GO:0048468	**cell development**	**GO:0019367**	**fatty acid elongation, saturated fatty acid**
GO:0006260	**DNA replication**	**GO:0071218**	**cellular response to misfolded protein**
GO:0051301	**cell division**	GO:0051788	response to misfolded protein
GO:0051276	**chromosome organization**	GO:0071630	nuclear protein quality control by the ubiquitin-proteasome system
GO:0071103	DNA conformation change	**GO:0009262**	**deoxyribonucleotide metabolic process**
GO:0006323	DNA packaging	**GO:0030497**	**fatty acid elongation**
GO:0031144	**proteasome localization**	**GO:0016926**	**protein desumoylation**
GO:0046328	**regulation of JNK cascade**	**GO:0031399**	**regulation of protein modification process**
GO:0034243	**regulation of transcription elongation from RNA polymerase II promoter**	**GO:0022613**	**ribonucleoprotein complex biogenesis**
GO:0070646	**protein modification by small protein removal**	**GO:0030261**	**chromosome condensation**
GO:0051783	**regulation of nuclear division**	**GO:0006259**	**DNA metabolic process**
GO:0033044	regulation of chromosome organization	**GO:0071824**	**protein-DNA complex subunit organization**

GO:0010639	negative regulation of organelle organization	**GO:0000723**	**telomere maintenance**
GO:0051983	**regulation of chromosome segregation**	**GO:0034501**	**protein localization to kinetochore**
GO:0031503	**protein-containing complex localization**	GO:0071459	protein localization to chromosome, centromeric region
GO:0051302	**regulation of cell division**	**GO:0009263**	**deoxyribonucleotide biosynthetic process**
GO:0010564	**regulation of cell cycle process**	**GO:0006352**	**DNA-templated transcription, initiation**
GO:0032954	regulation of cytokinetic process	**GO:0032200**	**telomere organization**
GO:0033047	regulation of mitotic sister chromatid segregation	**GO:0035803**	**egg coat formation**
GO:0032465	regulation of cytokinesis	**GO:0016071**	**mRNA metabolic process**
GO:0007088	regulation of mitotic nuclear division	**GO:0070601**	**centromeric sister chromatid cohesion**
GO:0090068	positive regulation of cell cycle process	**GO:0036297**	**interstrand cross-link repair**
GO:0045787	positive regulation of cell cycle	**GO:0048285**	**organelle fission**
GO:0000075	cell cycle checkpoint signaling	**GO:0000226**	**microtubule cytoskeleton organization**
GO:0010389	regulation of G2/M transition of mitotic cell cycle	GO:0007051	spindle organization
GO:1902749	regulation of cell cycle G2/M phase transition	GO:0031023	microtubule organizing center organization
GO:0045786	negative regulation of cell cycle	GO:0007020	microtubule nucleation
GO:0007346	regulation of mitotic cell cycle	GO:0051225	spindle assembly
GO:1901987	regulation of cell cycle phase transition	GO:0031109	microtubule polymerization or depolymerization
GO:2001251	negative regulation of chromosome organization	GO:0007052	mitotic spindle organization
GO:0031577	spindle checkpoint signaling	GO:0046785	microtubule polymerization
GO:0051985	negative regulation of chromosome segregation	**GO:0007339**	**binding of sperm to zona pellucida**
GO:1902100	negative regulation of metaphase/anaphase transition of cell cycle	GO:0009566	fertilization
GO:1901990	regulation of mitotic cell cycle phase transition	GO:0009988	cell-cell recognition
GO:0045930	negative regulation of mitotic cell cycle	GO:0007338	single fertilization
GO:0007094	mitotic spindle assembly checkpoint signaling	GO:0035036	sperm-egg recognition
GO:0000077	DNA damage checkpoint signaling	**GO:0051304**	**chromosome separation**
GO:0007093	mitotic cell cycle checkpoint signaling	**GO:0061640**	**cytoskeleton-dependent cytokinesis**
GO:0010948	negative regulation of cell cycle process	**GO:0009451**	**RNA modification**
GO:0031570	DNA integrity checkpoint signaling	**GO:0044770**	**cell cycle phase transition**
GO:0051784	negative regulation of nuclear division	**GO:2000431**	**regulation of cytokinesis, actomyosin contractile ring assembly**
GO:1901991	negative regulation of mitotic cell cycle phase transition	**GO:0006367**	**transcription initiation from RNA polymerase II promoter**
GO:0030071	regulation of mitotic metaphase/anaphase transition	**GO:0051169**	**nuclear transport**
GO:1901988	negative regulation of cell cycle phase transition	**GO:0034470**	**ncRNA processing**
GO:0045839	negative regulation of mitotic nuclear division	GO:0008380	RNA splicing
GO:1905818	regulation of chromosome separation	GO:0006397	mRNA processing
GO:0033045	regulation of sister chromatid segregation	GO:0008033	tRNA processing
GO:1902099	regulation of metaphase/anaphase transition of cell cycle	GO:0006402	mRNA catabolic process
GO:0010965	regulation of mitotic sister chromatid separation	GO:0000398	mRNA splicing, via spliceosome
GO:0045841	negative regulation of mitotic metaphase/anaphase transition	GO:0000956	nuclear-transcribed mRNA catabolic process
GO:0071173	spindle assembly checkpoint signaling	GO:0000375	RNA splicing, via transesterification reactions
GO:0071174	mitotic spindle checkpoint signaling	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:1905819	negative regulation of chromosome separation	**GO:0070925**	**organelle assembly**
GO:0033046	negative regulation of sister chromatid segregation	**GO:0022618**	**ribonucleoprotein complex assembly**
GO:2000816	negative regulation of mitotic sister chromatid separation	GO:0071826	ribonucleoprotein complex subunit organization
GO:0033048	negative regulation of mitotic sister chromatid segregation	GO:0000387	spliceosomal snRNP assembly
GO:0008299	**isoprenoid biosynthetic process**	**GO:0051129**	**negative regulation of cellular component organization**
GO:0016102	diterpenoid biosynthetic process	**GO:0006270**	**DNA replication initiation**
GO:0006776	vitamin A metabolic process	GO:0006269	DNA replication, synthesis of RNA primer
GO:0002138	retinoic acid biosynthetic process	**GO:0006401**	**RNA catabolic process**
GO:0042573	retinoic acid metabolic process	GO:0016075	rRNA catabolic process
GO:0016114	terpenoid biosynthetic process	GO:0034661	ncRNA catabolic process
GO:0000724	**double-strand break repair via homologous recombination**		

Supplementary Table A7. ***Drosophila melanogaster*** **germline-related enriched GO terms.** GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0000003	**reproduction**	**GO:0043543**	**protein acylation**
GO:0006260	**DNA replication**	**GO:0006289**	**nucleotide-excision repair**
GO:0006950	**response to stress**	**GO:0042158**	**lipoprotein biosynthetic process**
GO:0048646	**anatomical structure formation involved in morphogenesis**	GO:0006497	protein lipidation
GO:0051726	**regulation of cell cycle**	**GO:0010608**	**posttranscriptional regulation of gene expression**
GO:0022412	**cellular process involved in reproduction in multicellular organism**	**GO:0006261**	**DNA-dependent DNA replication**
GO:0030707	ovarian follicle cell development	**GO:0097435**	**supramolecular fiber organization**
GO:0007304	chorion-containing eggshell formation	**GO:0042254**	**ribosome biogenesis**
GO:0030703	eggshell formation	GO:0022613	ribonucleoprotein complex biogenesis
GO:0002066	columnar/cuboidal epithelial cell development	GO:0022618	ribonucleoprotein complex assembly
GO:0022414	reproductive process	GO:0006364	rRNA processing
GO:0032504	multicellular organism reproduction	**GO:0018193**	**peptidyl-amino acid modification**
GO:0007292	female gamete generation	**GO:0008213**	**protein alkylation**
GO:0044703	multi-organism reproductive process	**GO:0018205**	**peptidyl-lysine modification**
GO:0007281	germ cell development	**GO:0031109**	**microtubule polymerization or depolymerization**
GO:0003006	developmental process involved in reproduction	GO:0007020	microtubule nucleation
GO:0048477	oogenesis	GO:0046785	microtubule polymerization
GO:0019953	sexual reproduction	**GO:0043414**	**macromolecule methylation**
GO:0007276	gamete generation	**GO:0009263**	**deoxyribonucleotide biosynthetic process**
GO:0048609	multicellular organismal reproductive process	**GO:0016567**	**protein ubiquitination**
GO:0006913	**nucleocytoplasmic transport**	GO:0032446	protein modification by small protein conjugation
GO:0007059	**chromosome segregation**	**GO:0051258**	**protein polymerization**
GO:0006325	**chromatin organization**	**GO:0000280**	**nuclear division**
GO:0071103	DNA conformation change	GO:0051306	mitotic sister chromatid separation
GO:0022402	**cell cycle process**	GO:0140014	mitotic nuclear division
GO:1903047	mitotic cell cycle process	**GO:0048285**	**organelle fission**
GO:0000278	mitotic cell cycle	**GO:0018195**	**peptidyl-arginine modification**
GO:0000819	sister chromatid segregation	**GO:0006310**	**DNA recombination**
GO:0000910	cytokinesis	**GO:0000724**	**double-strand break repair via homologous recombination**
GO:0044772	mitotic cell cycle phase transition	**GO:0016573**	**histone acetylation**
GO:0007091	metaphase/anaphase transition of mitotic cell cycle	GO:0006473	protein acetylation
GO:0044784	metaphase/anaphase transition of cell cycle	GO:0006475	internal protein amino acid acetylation
GO:0098813	nuclear chromosome segregation	GO:0018394	peptidyl-lysine acetylation
GO:0000070	mitotic sister chromatid segregation	GO:0018393	internal peptidyl-lysine acetylation
GO:0051301	**cell division**	**GO:0006298**	**mismatch repair**
GO:0007049	**cell cycle**	**GO:0000725**	**recombinational repair**
GO:0032259	**methylation**	**GO:0030855**	**epithelial cell differentiation**
GO:0016072	**rRNA metabolic process**	GO:0002065	columnar/cuboidal epithelial cell differentiation
GO:0034470	ncRNA processing	GO:0002064	epithelial cell development
GO:0060249	**anatomical structure homeostasis**	**GO:0006302**	**double-strand break repair**
GO:0051983	**regulation of chromosome segregation**	**GO:0031297**	**replication fork processing**
GO:0016570	**histone modification**	**GO:0045786**	**negative regulation of cell cycle**
GO:0010629	**negative regulation of gene expression**	GO:0007088	regulation of mitotic nuclear division
GO:0006402	mRNA catabolic process	GO:0000075	cell cycle checkpoint signaling
GO:0010605	negative regulation of macromolecule metabolic process	GO:0010564	regulation of cell cycle process
GO:0000956	nuclear-transcribed mRNA catabolic process	GO:0007346	regulation of mitotic cell cycle
GO:0051253	negative regulation of RNA metabolic process	GO:1901987	regulation of cell cycle phase transition
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	GO:0031577	spindle checkpoint signaling
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	GO:1901990	regulation of mitotic cell cycle phase transition
GO:0009890	negative regulation of biosynthetic process	GO:0045930	negative regulation of mitotic cell cycle

GO:0051172	negative regulation of nitrogen compound metabolic process	GO:0007094	mitotic spindle assembly checkpoint signaling
GO:0045892	negative regulation of transcription, DNA-templated	GO:0000077	DNA damage checkpoint signaling
GO:0010558	negative regulation of macromolecule biosynthetic process	GO:0010948	negative regulation of cell cycle process
GO:1902679	negative regulation of RNA biosynthetic process	GO:1901991	negative regulation of mitotic cell cycle phase transition
GO:0031327	negative regulation of cellular biosynthetic process	GO:0031570	DNA integrity checkpoint signaling
GO:1903507	negative regulation of nucleic acid-templated transcription	GO:0051784	negative regulation of nuclear division
GO:0046474	**glycerophospholipid biosynthetic process**	GO:0007093	mitotic cell cycle checkpoint signaling
GO:0033043	**regulation of organelle organization**	GO:1901988	negative regulation of cell cycle phase transition
GO:2001251	negative regulation of chromosome organization	GO:0045839	negative regulation of mitotic nuclear division
GO:0051783	regulation of nuclear division	GO:0045841	negative regulation of mitotic metaphase/anaphase transition
GO:0033044	regulation of chromosome organization	GO:0071173	spindle assembly checkpoint signaling
GO:0033047	regulation of mitotic sister chromatid segregation	GO:0071174	mitotic spindle checkpoint signaling
GO:0051985	negative regulation of chromosome segregation	GO:1905819	negative regulation of chromosome separation
GO:1902100	negative regulation of metaphase/anaphase transition of cell cycle	GO:0033046	negative regulation of sister chromatid segregation
GO:0030071	regulation of mitotic metaphase/anaphase transition	GO:2000816	negative regulation of mitotic sister chromatid separation
GO:1905818	regulation of chromosome separation	GO:0033048	negative regulation of mitotic sister chromatid segregation
GO:0033045	regulation of sister chromatid segregation	**GO:0051304**	**chromosome separation**
GO:1902099	regulation of metaphase/anaphase transition of cell cycle	**GO:0061640**	**cytoskeleton-dependent cytokinesis**
GO:0010965	regulation of mitotic sister chromatid separation	**GO:0044770**	**cell cycle phase transition**
GO:0006270	**DNA replication initiation**	**GO:0071826**	**ribonucleoprotein complex subunit organization**
GO:0045005	DNA-dependent DNA replication maintenance of fidelity	**GO:0006479**	**protein methylation**
GO:0006338	**chromatin remodeling**	GO:0001510	RNA methylation
GO:0040029	regulation of gene expression, epigenetic	GO:0018216	peptidyl-arginine methylation
GO:0006265	**DNA topological change**	GO:0034968	histone lysine methylation
GO:0000723	**telomere maintenance**	GO:0016571	histone methylation
GO:0032200	**telomere organization**	GO:0018022	peptidyl-lysine methylation
GO:0051169	**nuclear transport**		

Supplementary Table A8. ***Ephydatia fluviatilis*** **germline-related enriched GO terms.** GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0006417	**regulation of translation**	**GO:0043574**	**peroxisomal transport**
GO:0006448	regulation of translational elongation	**GO:0007064**	**mitotic sister chromatid cohesion**
GO:0022613	**ribonucleoprotein complex biogenesis**	**GO:1903509**	**liposaccharide metabolic process**
GO:0042254	ribosome biogenesis	**GO:0050658**	**RNA transport**
GO:0042273	ribosomal large subunit biogenesis	GO:0050657	nucleic acid transport
GO:0042274	ribosomal small subunit biogenesis	GO:0006405	RNA export from nucleus
GO:0006364	rRNA processing	GO:0051236	establishment of RNA localization
GO:0071806	**protein transmembrane transport**	**GO:0006403**	**RNA localization**
GO:0006457	**protein folding**	**GO:0018216**	**peptidyl-arginine methylation**
GO:0008033	**tRNA processing**	GO:0006479	protein methylation
GO:0016072	rRNA metabolic process	**GO:0009451**	**RNA modification**
GO:0006418	tRNA aminoacylation for protein translation	**GO:0016180**	**snRNA processing**
GO:0006400	tRNA modification	**GO:0009067**	**aspartate family amino acid biosynthetic process**
GO:0001510	RNA methylation	**GO:0031123**	**RNA 3'-end processing**
GO:0043039	tRNA aminoacylation	**GO:0006520**	**cellular amino acid metabolic process**
GO:0032259	**methylation**	**GO:0006433**	**prolyl-tRNA aminoacylation**
GO:0006664	**glycolipid metabolic process**	**GO:0034227**	**tRNA thio-modification**
GO:0006505	GPI anchor metabolic process	**GO:2000765**	**regulation of cytoplasmic translation**
GO:0009247	glycolipid biosynthetic process	**GO:0018202**	**peptidyl-histidine modification**
GO:0046467	membrane lipid biosynthetic process	**GO:0006643**	**membrane lipid metabolic process**

GO:0043038	**amino acid activation**	**GO:0006424**	**glutamyl-tRNA aminoacylation**
GO:0043414	**macromolecule methylation**	**GO:0017038**	**protein import**
GO:0042158	**lipoprotein biosynthetic process**	**GO:0065002**	**intracellular protein transmembrane transport**
GO:0006506	GPI anchor biosynthetic process	GO:0072594	establishment of protein localization to organelle
GO:0006497	protein lipidation	GO:0006625	protein targeting to peroxisome
GO:0051189	**prosthetic group metabolic process**	GO:0033365	protein localization to organelle
GO:0042157	**lipoprotein metabolic process**	GO:0006605	protein targeting
GO:0031503	**protein-containing complex localization**	GO:0044743	protein transmembrane import into intracellular organelle
GO:0006839	**mitochondrial transport**	GO:0006626	protein targeting to mitochondrion
GO:0045454	**cell redox homeostasis**	GO:0070585	protein localization to mitochondrion
GO:0006777	**Mo-molybdopterin cofactor biosynthetic process**	GO:0072662	protein localization to peroxisome
GO:0043545	molybdopterin cofactor metabolic process	GO:0030150	protein import into mitochondrial matrix
GO:0019720	Mo-molybdopterin cofactor metabolic process	GO:0016560	protein import into peroxisome matrix, docking
GO:0002182	**cytoplasmic translational elongation**	GO:0016558	protein import into peroxisome matrix
GO:1990542	**mitochondrial transmembrane transport**	GO:0015919	peroxisomal membrane transport
GO:0050684	**regulation of mRNA processing**	GO:0072655	establishment of protein localization to mitochondrion
GO:0048024	regulation of mRNA splicing, via spliceosome	GO:0072663	establishment of protein localization to peroxisome
GO:0000381	regulation of alternative mRNA splicing, via spliceosome	**GO:0044085**	**cellular component biogenesis**
GO:0043484	**regulation of RNA splicing**	**GO:0046474**	**glycerophospholipid biosynthetic process**
GO:0018195	**peptidyl-arginine modification**	GO:0008654	phospholipid biosynthetic process
GO:1903311	**regulation of mRNA metabolic process**	GO:0006661	phosphatidylinositol biosynthetic process
GO:0006766	**vitamin metabolic process**	GO:0045017	glycerolipid biosynthetic process
GO:0071166	**ribonucleoprotein complex localization**	**GO:0009066**	**aspartate family amino acid metabolic process**
GO:0043631	**RNA polyadenylation**	**GO:0051169**	**nuclear transport**
GO:0006414	**translational elongation**	**GO:0008380**	**RNA splicing**
GO:0006487	**protein N-linked glycosylation**	**GO:0002097**	**tRNA wobble base modification**
GO:0008213	**protein alkylation**	**GO:0006913**	**nucleocytoplasmic transport**
GO:0006260	**DNA replication**	GO:0006611	protein export from nucleus
GO:0006261	**DNA-dependent DNA replication**	**GO:0034622**	**cellular protein-containing complex assembly**
GO:0006281	**DNA repair**	GO:0033108	mitochondrial respiratory chain complex assembly
GO:0006289	nucleotide-excision repair	GO:0071824	protein-DNA complex subunit organization
GO:0006974	cellular response to DNA damage stimulus	GO:0065004	protein-DNA complex assembly
GO:0034248	**regulation of cellular amide metabolic process**	GO:0017004	cytochrome complex assembly
GO:0010608	**posttranscriptional regulation of gene expression**	GO:0071826	ribonucleoprotein complex subunit organization
GO:0000380	**alternative mRNA splicing, via spliceosome**	GO:0022618	ribonucleoprotein complex assembly
GO:0032543	**mitochondrial translation**	GO:0000387	spliceosomal snRNP assembly
GO:0007005	**mitochondrion organization**	GO:0034728	nucleosome organization
GO:0017182	**peptidyl-diphthamide metabolic process**	**GO:0017183**	**peptidyl-diphthamide biosynthetic process from peptidyl-histidine**
GO:0002181	**cytoplasmic translation**	GO:1900247	regulation of cytoplasmic translational elongation
GO:0001522	**pseudouridine synthesis**	**GO:0042176**	**regulation of protein catabolic process**
GO:0007031	**peroxisome organization**	**GO:0002098**	**tRNA wobble uridine modification**
GO:0006401	**RNA catabolic process**	**GO:0007020**	**microtubule nucleation**
GO:0018193	**peptidyl-amino acid modification**	GO:0046785	microtubule polymerization
GO:0006333	**chromatin assembly or disassembly**	**GO:0006397**	**mRNA processing**
GO:0031497	chromatin assembly	GO:0031124	mRNA 3'-end processing
GO:0006334	nucleosome assembly	GO:0006378	mRNA polyadenylation
GO:0016073	**snRNA metabolic process**	GO:0000398	mRNA splicing, via spliceosome
GO:0006413	**translational initiation**	GO:0000375	RNA splicing, via transesterification reactions
GO:0016071	**mRNA metabolic process**	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0015931	**nucleobase-containing compound transport**	**GO:0006270**	**DNA replication initiation**

Supplementary Table A9. *Haliotis rufescens* germline-related enriched GO terms. GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0006629	**lipid metabolic process**	**GO:1901136**	**carbohydrate derivative catabolic process**
GO:0031396	**regulation of protein ubiquitination**	**GO:0044242**	**cellular lipid catabolic process**
GO:1903320	regulation of protein modification by small protein conjugation or removal	GO:0030149	sphingolipid catabolic process
GO:0042254	**ribosome biogenesis**	GO:0046466	membrane lipid catabolic process
GO:0022613	ribonucleoprotein complex biogenesis	GO:0006687	glycosphingolipid metabolic process
GO:0006364	rRNA processing	GO:0006672	ceramide metabolic process
GO:0031589	**cell-substrate adhesion**	GO:0006689	ganglioside catabolic process
GO:0006457	**protein folding**	GO:0006685	sphingomyelin catabolic process
GO:0006665	**sphingolipid metabolic process**	GO:0046479	glycosphingolipid catabolic process
GO:0006664	glycolipid metabolic process	GO:0019377	glycolipid catabolic process
GO:0034470	**ncRNA processing**	GO:0046514	ceramide catabolic process
GO:0016072	rRNA metabolic process	**GO:1903509**	**liposaccharide metabolic process**
GO:0006260	**DNA replication**	**GO:0006643**	**membrane lipid metabolic process**
GO:0006270	**DNA replication initiation**	**GO:0001573**	**ganglioside metabolic process**
GO:0006323	**DNA packaging**	**GO:0007160**	**cell-matrix adhesion**
GO:0051225	**spindle assembly**	**GO:0030261**	**chromosome condensation**
GO:0007051	spindle organization	GO:0007076	mitotic chromosome condensation
GO:0016042	**lipid catabolic process**	**GO:0006684**	**sphingomyelin metabolic process**
GO:0034660	**ncRNA metabolic process**		

Supplementary Table A10. *Ruditapes philippinarum* germline-related enriched GO terms. GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0000003	**reproduction**	**GO:0000380**	**alternative mRNA splicing, via spliceosome**
GO:0006260	**DNA replication**	**GO:0034587**	**piRNA metabolic process**
GO:0051726	**regulation of cell cycle**	**GO:0006473**	**protein acetylation**
GO:0031589	**cell-substrate adhesion**	GO:0018393	internal peptidyl-lysine acetylation
GO:0006606	**protein import into nucleus**	GO:0006475	internal protein amino acid acetylation
GO:0034504	protein localization to nucleus	GO:0018394	peptidyl-lysine acetylation
GO:0051170	import into nucleus	**GO:0016570**	**histone modification**
GO:0006913	nucleocytoplasmic transport	**GO:0008213**	**protein alkylation**
GO:0022412	**cellular process involved in reproduction in multicellular organism**	**GO:0018205**	**peptidyl-lysine modification**
GO:0022414	reproductive process	**GO:0006261**	**DNA-dependent DNA replication**
GO:0007286	spermatid development	**GO:0051383**	**kinetochore organization**
GO:0051321	meiotic cell cycle	**GO:0006189**	**'de novo' IMP biosynthetic process**
GO:0032504	multicellular organism reproduction	GO:0009129	pyrimidine nucleoside monophosphate metabolic process
GO:0044703	multi-organism reproductive process	GO:0009130	pyrimidine nucleoside monophosphate biosynthetic process
GO:0007281	germ cell development	GO:0046040	IMP metabolic process
GO:0048515	spermatid differentiation	GO:0006188	IMP biosynthetic process
GO:0003006	developmental process involved in reproduction	**GO:0019692**	**deoxyribose phosphate metabolic process**
GO:0007283	spermatogenesis	**GO:0006424**	**glutamyl-tRNA aminoacylation**
GO:0007276	gamete generation	**GO:0006336**	**DNA replication-independent chromatin assembly**
GO:0048232	male gamete generation	GO:0034080	CENP-A containing chromatin assembly
GO:0019953	sexual reproduction	**GO:0018193**	**peptidyl-amino acid modification**
GO:0048609	multicellular organismal reproductive process	**GO:0009119**	**ribonucleoside metabolic process**
GO:0007059	**chromosome segregation**	**GO:0046112**	**nucleobase biosynthetic process**
GO:0007017	**microtubule-based process**	GO:0009113	purine nucleobase biosynthetic process

GO:0022402	**cell cycle process**	**GO:0044782**	**cilium organization**
GO:1903047	mitotic cell cycle process	GO:0030031	cell projection assembly
GO:0000278	mitotic cell cycle	GO:0060271	cilium assembly
GO:0000910	cytokinesis	GO:0120036	plasma membrane bounded cell projection organization
GO:0000281	mitotic cytokinesis	GO:0120031	plasma membrane bounded cell projection assembly
GO:0044772	mitotic cell cycle phase transition	GO:0035082	axoneme assembly
GO:1902850	microtubule cytoskeleton organization involved in mitosis	GO:0070286	axonemal dynein complex assembly
GO:0007091	metaphase/anaphase transition of mitotic cell cycle	**GO:0032365**	**intracellular lipid transport**
GO:0044784	metaphase/anaphase transition of cell cycle	**GO:0050953**	**sensory perception of light stimulus**
GO:0000086	G2/M transition of mitotic cell cycle	**GO:0071824**	**protein-DNA complex subunit organization**
GO:0044839	cell cycle G2/M phase transition	**GO:0000280**	**nuclear division**
GO:0006457	**protein folding**	GO:0051306	mitotic sister chromatid separation
GO:0051301	**cell division**	GO:0007080	mitotic metaphase plate congression
GO:0007049	**cell cycle**	GO:0007131	reciprocal meiotic recombination
GO:0006974	**cellular response to DNA damage stimulus**	GO:1903046	meiotic cell cycle process
GO:0033554	cellular response to stress	GO:0140013	meiotic nuclear division
GO:0006281	DNA repair	GO:0140014	mitotic nuclear division
GO:0006996	**organelle organization**	GO:0051315	attachment of mitotic spindle microtubules to kinetochore
GO:0032259	**methylation**	GO:0051310	metaphase plate congression
GO:0051438	**regulation of ubiquitin-protein transferase activity**	**GO:0061641**	**CENP-A containing chromatin organization**
GO:0003352	**regulation of cilium movement**	**GO:0048285**	**organelle fission**
GO:0060632	regulation of microtubule-based movement	**GO:0031055**	**chromatin remodeling at centromere**
GO:0006396	**RNA processing**	**GO:0000226**	**microtubule cytoskeleton organization**
GO:0034660	ncRNA metabolic process	GO:0001578	microtubule bundle formation
GO:0051983	**regulation of chromosome segregation**	GO:0007098	centrosome cycle
GO:0051302	**regulation of cell division**	GO:0003341	cilium movement
GO:0032886	**regulation of microtubule-based process**	GO:0031023	microtubule organizing center organization
GO:0033044	**regulation of chromosome organization**	GO:0051225	spindle assembly
GO:2001251	negative regulation of chromosome organization	GO:0007051	spindle organization
GO:0051783	regulation of nuclear division	GO:0007018	microtubule-based movement
GO:0033047	regulation of mitotic sister chromatid segregation	GO:0051298	centrosome duplication
GO:0051985	negative regulation of chromosome segregation	**GO:0017038**	**protein import**
GO:1902100	negative regulation of metaphase/anaphase transition of cell cycle	**GO:0006259**	**DNA metabolic process**
GO:0030071	regulation of mitotic metaphase/anaphase transition	**GO:0034508**	**centromere complex assembly**
GO:1905818	regulation of chromosome separation	GO:0043486	histone exchange
GO:0033045	regulation of sister chromatid segregation	GO:0051382	kinetochore assembly
GO:1902099	regulation of metaphase/anaphase transition of cell cycle	GO:0065004	protein-DNA complex assembly
GO:0010965	regulation of mitotic sister chromatid separation	GO:0034728	nucleosome organization
GO:0051052	**regulation of DNA metabolic process**	**GO:0034724**	**DNA replication-independent chromatin organization**
GO:0007346	**regulation of mitotic cell cycle**	**GO:0030030**	**cell projection organization**
GO:0032465	regulation of cytokinesis	**GO:0006325**	**chromatin organization**
GO:0007088	regulation of mitotic nuclear division	GO:0071103	DNA conformation change
GO:0000075	cell cycle checkpoint signaling	**GO:0042254**	**ribosome biogenesis**
GO:0045786	negative regulation of cell cycle	GO:0022613	ribonucleoprotein complex biogenesis
GO:0010564	regulation of cell cycle process	GO:0006364	rRNA processing
GO:1901987	regulation of cell cycle phase transition	**GO:0016571**	**histone methylation**
GO:0031577	spindle checkpoint signaling	GO:0034968	histone lysine methylation
GO:1901990	regulation of mitotic cell cycle phase transition	GO:0018023	peptidyl-lysine trimethylation
GO:0045930	negative regulation of mitotic cell cycle	GO:0006479	protein methylation
GO:0007094	mitotic spindle assembly checkpoint signaling	GO:0018022	peptidyl-lysine methylation
GO:0010948	negative regulation of cell cycle process	**GO:0051169**	**nuclear transport**
GO:1901991	negative regulation of mitotic cell cycle phase transition	**GO:0051304**	**chromosome separation**
GO:0031570	DNA integrity checkpoint signaling	**GO:0000288**	**nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay**
GO:0051784	negative regulation of nuclear division	**GO:0031399**	**regulation of protein modification process**
GO:0007093	mitotic cell cycle checkpoint signaling	**GO:0061640**	**cytoskeleton-dependent cytokinesis**
GO:1901988	negative regulation of cell cycle phase transition	**GO:0009451**	**RNA modification**
GO:0045839	negative regulation of mitotic nuclear division	**GO:0044770**	**cell cycle phase transition**

GO:0045841	negative regulation of mitotic metaphase/anaphase transition	**GO:0000724**	**double-strand break repair via homologous recombination**
GO:0071173	spindle assembly checkpoint signaling	**GO:0060491**	**regulation of cell projection assembly**
GO:0071174	mitotic spindle checkpoint signaling	GO:0120032	regulation of plasma membrane bounded cell projection assembly
GO:1905819	negative regulation of chromosome separation	GO:1902017	regulation of cilium assembly
GO:0033046	negative regulation of sister chromatid segregation	**GO:0000725**	**recombinational repair**
GO:2000816	negative regulation of mitotic sister chromatid separation	**GO:0007160**	**cell-matrix adhesion**
GO:0033048	negative regulation of mitotic sister chromatid segregation	**GO:0051276**	**chromosome organization**
GO:0043414	**macromolecule methylation**	**GO:0034470**	**ncRNA processing**
GO:0001510	RNA methylation	GO:0016072	rRNA metabolic process
GO:0006400	tRNA modification	GO:0008033	tRNA processing
GO:0090329	**regulation of DNA-dependent DNA replication**	**GO:0015918**	**sterol transport**
GO:2000104	negative regulation of DNA-dependent DNA replication	GO:0032366	intracellular sterol transport
GO:0048478	replication fork protection	GO:0030301	cholesterol transport
GO:0006275	**regulation of DNA replication**	GO:0032367	intracellular cholesterol transport
GO:0051053	**negative regulation of DNA metabolic process**	**GO:0070925**	**organelle assembly**
GO:2000779	regulation of double-strand break repair	**GO:0006302**	**double-strand break repair**
GO:2000780	negative regulation of double-strand break repair	**GO:0008608**	**attachment of spindle microtubules to kinetochore**
GO:2001020	regulation of response to DNA damage stimulus	GO:0000819	sister chromatid segregation
GO:2001021	negative regulation of response to DNA damage stimulus	GO:0098813	nuclear chromosome segregation
GO:0080135	regulation of cellular response to stress	GO:0000070	mitotic sister chromatid segregation
GO:0000018	regulation of DNA recombination	**GO:0006270**	**DNA replication initiation**
GO:0045910	negative regulation of DNA recombination	GO:0045005	DNA-dependent DNA replication maintenance of fidelity
GO:0006282	regulation of DNA repair	GO:0006269	DNA replication, synthesis of RNA primer
GO:0010569	regulation of double-strand break repair via homologous recombination	**GO:0006338**	**chromatin remodeling**
GO:0045738	negative regulation of DNA repair	**GO:0006289**	**nucleotide-excision repair**
GO:2000042	negative regulation of double-strand break repair via homologous recombination	**GO:0072527**	**pyrimidine-containing compound metabolic process**
GO:0072528	**pyrimidine-containing compound biosynthetic process**	**GO:0035825**	**homologous recombination**
GO:0046131	pyrimidine ribonucleoside metabolic process	**GO:0031401**	**positive regulation of protein modification process**
GO:0006213	pyrimidine nucleoside metabolic process	GO:0051247	positive regulation of protein metabolic process
GO:0009221	pyrimidine deoxyribonucleotide biosynthetic process	GO:0031396	regulation of protein ubiquitination
GO:0009263	deoxyribonucleotide biosynthetic process	GO:1904668	positive regulation of ubiquitin protein ligase activity
GO:0006221	pyrimidine nucleotide biosynthetic process	GO:0032270	positive regulation of cellular protein metabolic process
GO:0006220	pyrimidine nucleotide metabolic process	GO:0031398	positive regulation of protein ubiquitination
GO:0046385	deoxyribose phosphate biosynthetic process	GO:1904666	regulation of ubiquitin protein ligase activity
GO:0009394	2'-deoxyribonucleotide metabolic process	GO:0051443	positive regulation of ubiquitin-protein transferase activity
GO:0009219	pyrimidine deoxyribonucleotide metabolic process		
GO:0009265	2'-deoxyribonucleotide biosynthetic process		

Supplementary Table A11. ***Schmidtea mediterranea*** **germline-related enriched GO terms.** GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0000003	**reproduction**	**GO:0016570**	**histone modification**
GO:0022414	**reproductive process**	**GO:0008213**	**protein alkylation**
GO:0051321	meiotic cell cycle	**GO:0006261**	**DNA-dependent DNA replication**
GO:0033554	**cellular response to stress**	**GO:1901571**	**fatty acid derivative transport**
GO:0051246	**regulation of protein metabolic process**	**GO:0070525**	**tRNA threonylcarbamoyladenosine metabolic process**
GO:0006913	**nucleocytoplasmic transport**	**GO:0006310**	**DNA recombination**
GO:0006611	protein export from nucleus	**GO:0050482**	**arachidonic acid secretion**
GO:0006606	protein import into nucleus	GO:0015908	fatty acid transport

GO:0051170	import into nucleus	GO:0015909	long-chain fatty acid transport
GO:0051168	nuclear export	GO:1903963	arachidonate transport
GO:0007059	**chromosome segregation**	GO:0032309	icosanoid secretion
GO:0022402	**cell cycle process**	**GO:0071826**	**ribonucleoprotein complex subunit organization**
GO:1903047	mitotic cell cycle process	**GO:0016925**	**protein sumoylation**
GO:0000278	mitotic cell cycle	**GO:0035825**	**homologous recombination**
GO:1902969	mitotic DNA replication	**GO:0015931**	**nucleobase-containing compound transport**
GO:0000819	sister chromatid segregation	**GO:0018193**	**peptidyl-amino acid modification**
GO:0007076	mitotic chromosome condensation	**GO:0001522**	**pseudouridine synthesis**
GO:0098813	nuclear chromosome segregation	**GO:0046112**	**nucleobase biosynthetic process**
GO:1902298	cell cycle DNA replication maintenance of fidelity	GO:0009113	purine nucleobase biosynthetic process
GO:1990426	mitotic recombination-dependent replication fork processing	GO:0006144	purine nucleobase metabolic process
GO:0000070	mitotic sister chromatid segregation	**GO:0006403**	**RNA localization**
GO:0007062	sister chromatid cohesion	**GO:0006950**	**response to stress**
GO:0033260	nuclear DNA replication	**GO:0017038**	**protein import**
GO:1990505	mitotic DNA replication maintenance of fidelity	**GO:0016180**	**snRNA processing**
GO:0006260	**DNA replication**	**GO:0036159**	**inner dynein arm assembly**
GO:0007049	**cell cycle**	**GO:0016073**	**snRNA metabolic process**
GO:0022613	**ribonucleoprotein complex biogenesis**	**GO:0051276**	**chromosome organization**
GO:0042254	ribosome biogenesis	GO:0071103	DNA conformation change
GO:0042255	ribosome assembly	GO:0006323	DNA packaging
GO:0042273	ribosomal large subunit biogenesis	GO:0006265	DNA topological change
GO:0022618	ribonucleoprotein complex assembly	GO:0006325	chromatin organization
GO:0000154	rRNA modification	**GO:0006352**	**DNA-templated transcription, initiation**
GO:0000245	spliceosomal complex assembly	**GO:0030261**	**chromosome condensation**
GO:0000387	spliceosomal snRNP assembly	GO:0031497	chromatin assembly
GO:0000027	ribosomal large subunit assembly	GO:0006334	nucleosome assembly
GO:0006364	rRNA processing	**GO:0018205**	**peptidyl-lysine modification**
GO:0032259	**methylation**	**GO:0046717**	**acid secretion**
GO:0034470	**ncRNA processing**	**GO:0044839**	**cell cycle G2/M phase transition**
GO:0008380	RNA splicing	**GO:0016071**	**mRNA metabolic process**
GO:0006397	mRNA processing	**GO:0034504**	**protein localization to nucleus**
GO:0016072	rRNA metabolic process	**GO:0000725**	**recombinational repair**
GO:0008033	tRNA processing	**GO:0006298**	**mismatch repair**
GO:0006400	tRNA modification	**GO:0000966**	**RNA 5'-end processing**
GO:0001510	RNA methylation	**GO:0034471**	**ncRNA 5'-end processing**
GO:0006399	tRNA metabolic process	GO:0001682	tRNA 5'-leader removal
GO:0000398	mRNA splicing, via spliceosome	GO:0099116	tRNA 5'-end processing
GO:0000375	RNA splicing, via transesterification reactions	**GO:0006302**	**double-strand break repair**
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	**GO:0000724**	**double-strand break repair via homologous recombination**
GO:0008295	**spermidine biosynthetic process**	**GO:0000280**	**nuclear division**
GO:0008215	spermine metabolic process	GO:0007131	reciprocal meiotic recombination
GO:0006597	spermine biosynthetic process	GO:0140013	meiotic nuclear division
GO:0008216	spermidine metabolic process	GO:0007127	meiosis I
GO:0031503	**protein-containing complex localization**	GO:1903046	meiotic cell cycle process
GO:0050658	**RNA transport**	GO:0140014	mitotic nuclear division
GO:0050657	nucleic acid transport	GO:0061982	meiosis I cell cycle process
GO:0006406	mRNA export from nucleus	**GO:0006289**	**nucleotide-excision repair**
GO:0006405	RNA export from nucleus	**GO:0009451**	**RNA modification**
GO:0051028	mRNA transport	**GO:0032270**	**positive regulation of cellular protein metabolic process**
GO:0051236	establishment of RNA localization	GO:0051247	positive regulation of protein metabolic process
GO:0043414	**macromolecule methylation**	GO:0031401	positive regulation of protein modification process
GO:0072528	**pyrimidine-containing compound biosynthetic process**	**GO:0048285**	**organelle fission**
GO:0072527	**pyrimidine-containing compound metabolic process**	**GO:0006284**	**base-excision repair**
GO:0016572	**histone phosphorylation**	**GO:0002949**	**tRNA threonylcarbamoyladenosine modification**
GO:0009112	**nucleobase metabolic process**	**GO:0045934**	**negative regulation of nucleobase-containing compound metabolic process**
GO:0018195	**peptidyl-arginine modification**	**GO:0034728**	**nucleosome organization**

GO:0042176	**regulation of protein catabolic process**	GO:0006333	chromatin assembly or disassembly
GO:0009894	**regulation of catabolic process**	GO:0065004	protein-DNA complex assembly
GO:0000075	**cell cycle checkpoint signaling**	**GO:0006354**	**DNA-templated transcription, elongation**
GO:0045786	negative regulation of cell cycle	**GO:0006367**	**transcription initiation from RNA polymerase II promoter**
GO:0010389	regulation of G2/M transition of mitotic cell cycle	GO:0006368	transcription elongation from RNA polymerase II promoter
GO:1902749	regulation of cell cycle G2/M phase transition	**GO:0034660**	**ncRNA metabolic process**
GO:0044773	mitotic DNA damage checkpoint signaling	**GO:0051169**	**nuclear transport**
GO:0044774	mitotic DNA integrity checkpoint signaling	**GO:0071824**	**protein-DNA complex subunit organization**
GO:0007095	mitotic G2 DNA damage checkpoint signaling	**GO:0006479**	**protein methylation**
GO:0000077	DNA damage checkpoint signaling	GO:0018216	peptidyl-arginine methylation
GO:0007093	mitotic cell cycle checkpoint signaling	GO:0035246	peptidyl-arginine N-methylation
GO:0044818	mitotic G2/M transition checkpoint	**GO:0051053**	**negative regulation of DNA metabolic process**
GO:0031570	DNA integrity checkpoint signaling	**GO:0071715**	**icosanoid transport**
GO:0071166	**ribonucleoprotein complex localization**	GO:0015718	monocarboxylic acid transport
GO:0044786	**cell cycle DNA replication**	**GO:0006270**	**DNA replication initiation**

Supplementary Table A12. ***Xenopus tropicalis*** **germline-related enriched GO terms.** GO terms were collapsed for semantic similarity (SimRel cut-off of 0.7). Representative GO terms are highlighted in bold and in grey rows: all following GO terms (until the subsequent grey bold row) belong to that semantic group.

TermID	Name	TermID	Name
GO:0003006	**developmental process involved in reproduction**	**GO:0034248**	**regulation of cellular amide metabolic process**
GO:0006281	**DNA repair**	**GO:0043543**	**protein acylation**
GO:0000725	recombinational repair	**GO:0006261**	**DNA-dependent DNA replication**
GO:0006302	double-strand break repair	**GO:0071824**	**protein-DNA complex subunit organization**
GO:0006289	nucleotide-excision repair	**GO:0016570**	**histone modification**
GO:0006284	base-excision repair	**GO:0008213**	**protein alkylation**
GO:0033554	cellular response to stress	**GO:0018205**	**peptidyl-lysine modification**
GO:0051726	**regulation of cell cycle**	**GO:0043414**	**macromolecule methylation**
GO:0030010	**establishment of cell polarity**	GO:0001510	RNA methylation
GO:0006913	**nucleocytoplasmic transport**	GO:0001522	pseudouridine synthesis
GO:0006611	protein export from nucleus	GO:0030488	tRNA methylation
GO:0051168	nuclear export	GO:0006479	protein methylation
GO:0007059	**chromosome segregation**	GO:0006400	tRNA modification
GO:0022402	**cell cycle process**	**GO:0034470**	**ncRNA processing**
GO:1903047	mitotic cell cycle process	GO:0008380	RNA splicing
GO:0000278	mitotic cell cycle	GO:0006397	mRNA processing
GO:0000819	sister chromatid segregation	GO:0016072	rRNA metabolic process
GO:0000910	cytokinesis	GO:0008033	tRNA processing
GO:0000281	mitotic cytokinesis	GO:0006402	mRNA catabolic process
GO:0044772	mitotic cell cycle phase transition	GO:0006399	tRNA metabolic process
GO:1902850	microtubule cytoskeleton organization involved in mitosis	GO:0000398	mRNA splicing, via spliceosome
GO:0008608	attachment of spindle microtubules to kinetochore	GO:0000956	nuclear-transcribed mRNA catabolic process
GO:0007091	metaphase/anaphase transition of mitotic cell cycle	GO:0000375	RNA splicing, via transesterification reactions
GO:0044784	metaphase/anaphase transition of cell cycle	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0098813	nuclear chromosome segregation	**GO:0006325**	**chromatin organization**
GO:0000086	G2/M transition of mitotic cell cycle	GO:0071103	DNA conformation change
GO:0044839	cell cycle G2/M phase transition	**GO:0043488**	**regulation of mRNA stability**
GO:0000070	mitotic sister chromatid segregation	GO:0043487	regulation of RNA stability
GO:0007062	sister chromatid cohesion	GO:0000290	deadenylation-dependent decapping of nuclear-transcribed mRNA
GO:0051301	**cell division**	GO:0061013	regulation of mRNA catabolic process
GO:0022613	**ribonucleoprotein complex biogenesis**	GO:0000288	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
GO:0042254	ribosome biogenesis	**GO:0018095**	**protein polyglutamylation**

GO:0042274	ribosomal small subunit biogenesis	**GO:0032446**	**protein modification by small protein conjugation**
GO:0022618	ribonucleoprotein complex assembly	GO:0016579	protein deubiquitination
GO:0000245	spliceosomal complex assembly	GO:0070646	protein modification by small protein removal
GO:0000387	spliceosomal snRNP assembly	GO:0016567	protein ubiquitination
GO:0006364	rRNA processing	**GO:0006403**	**RNA localization**
GO:0032259	**methylation**	**GO:0006401**	**RNA catabolic process**
GO:0060249	**anatomical structure homeostasis**	**GO:0018200**	**peptidyl-glutamic acid modification**
GO:0051983	**regulation of chromosome segregation**	**GO:0006259**	**DNA metabolic process**
GO:0031503	**protein-containing complex localization**	**GO:0016180**	**snRNA processing**
GO:0051302	**regulation of cell division**	GO:0034472	snRNA 3'-end processing
GO:0033044	**regulation of chromosome organization**	**GO:0016073**	**snRNA metabolic process**
GO:2001251	negative regulation of chromosome organization	**GO:0070601**	**centromeric sister chromatid cohesion**
GO:0051783	regulation of nuclear division	**GO:0000280**	**nuclear division**
GO:0051052	**regulation of DNA metabolic process**	GO:0051306	mitotic sister chromatid separation
GO:0015693	**magnesium ion transport**	GO:0140014	mitotic nuclear division
GO:0050658	**RNA transport**	**GO:0018193**	**peptidyl-amino acid modification**
GO:0050657	nucleic acid transport	**GO:0006352**	**DNA-templated transcription, initiation**
GO:0006405	RNA export from nucleus	**GO:0048285**	**organelle fission**
GO:0051236	establishment of RNA localization	**GO:0090501**	**RNA phosphodiester bond hydrolysis**
GO:0010564	**regulation of cell cycle process**	**GO:0000226**	**microtubule cytoskeleton organization**
GO:0032465	regulation of cytokinesis	GO:0042073	intraciliary transport
GO:0007088	regulation of mitotic nuclear division	GO:0007098	centrosome cycle
GO:0090068	positive regulation of cell cycle process	GO:0007051	spindle organization
GO:0045787	positive regulation of cell cycle	GO:0031023	microtubule organizing center organization
GO:0000075	cell cycle checkpoint signaling	GO:0007020	microtubule nucleation
GO:1902749	regulation of cell cycle G2/M phase transition	GO:0051225	spindle assembly
GO:0032467	positive regulation of cytokinesis	GO:0007052	mitotic spindle organization
GO:0045786	negative regulation of cell cycle	GO:0007099	centriole replication
GO:0007346	regulation of mitotic cell cycle	GO:0098534	centriole assembly
GO:1901987	regulation of cell cycle phase transition	GO:0051298	centrosome duplication
GO:0051781	positive regulation of cell division	**GO:0045292**	**mRNA cis splicing, via spliceosome**
GO:0031577	spindle checkpoint signaling	**GO:0016071**	**mRNA metabolic process**
GO:1902100	negative regulation of metaphase/anaphase transition of cell cycle	**GO:0051383**	**kinetochore organization**
GO:0051985	negative regulation of chromosome segregation	**GO:1903320**	**regulation of protein modification by small protein conjugation or removal**
GO:0033047	regulation of mitotic sister chromatid segregation	**GO:0030490**	**maturation of SSU-rRNA**
GO:1901990	regulation of mitotic cell cycle phase transition	**GO:0043628**	**ncRNA 3'-end processing**
GO:0045930	negative regulation of mitotic cell cycle	GO:0031124	mRNA 3'-end processing
GO:0007094	mitotic spindle assembly checkpoint signaling	**GO:0051656**	**establishment of organelle localization**
GO:0007093	mitotic cell cycle checkpoint signaling	**GO:0031297**	**replication fork processing**
GO:0010948	negative regulation of cell cycle process	**GO:0036297**	**interstrand cross-link repair**
GO:0051784	negative regulation of nuclear division	**GO:0009451**	**RNA modification**
GO:1901991	negative regulation of mitotic cell cycle phase transition	**GO:0031123**	**RNA 3'-end processing**
GO:0030071	regulation of mitotic metaphase/anaphase transition	**GO:0070925**	**organelle assembly**
GO:1901988	negative regulation of cell cycle phase transition	**GO:0030261**	**chromosome condensation**
GO:0045839	negative regulation of mitotic nuclear division	**GO:0051304**	**chromosome separation**
GO:1905818	regulation of chromosome separation	**GO:0061640**	**cytoskeleton-dependent cytokinesis**
GO:0033045	regulation of sister chromatid segregation	**GO:0065004**	**protein-DNA complex assembly**
GO:1902099	regulation of metaphase/anaphase transition of cell cycle	GO:0051382	kinetochore assembly
GO:0010965	regulation of mitotic sister chromatid separation	GO:0034508	centromere complex assembly
GO:0045841	negative regulation of mitotic metaphase/anaphase transition	GO:0070897	transcription preinitiation complex assembly
GO:0071173	spindle assembly checkpoint signaling	**GO:0044770**	**cell cycle phase transition**
GO:0071174	mitotic spindle checkpoint signaling	**GO:0010390**	**histone monoubiquitination**
GO:1905819	negative regulation of chromosome separation	**GO:0006383**	**transcription by RNA polymerase III**
GO:0033046	negative regulation of sister chromatid segregation	**GO:0000413**	**protein peptidyl-prolyl isomerization**
GO:2000816	negative regulation of mitotic sister chromatid separation	**GO:0018208**	**peptidyl-proline modification**

GO:0033048	negative regulation of mitotic sister chromatid segregation	**GO:0006354**	**DNA-templated transcription, elongation**
GO:0070647	**protein modification by small protein conjugation or removal**	**GO:0000723**	**telomere maintenance**
GO:0006260	**DNA replication**	GO:0016233	telomere capping
GO:0008654	**phospholipid biosynthetic process**	**GO:0016574**	**histone ubiquitination**
GO:0007173	**epidermal growth factor receptor signaling pathway**	**GO:0006367**	**transcription initiation from RNA polymerase II promoter**
GO:0038127	ERBB signaling pathway	GO:0006368	transcription elongation from RNA polymerase II promoter
GO:0043631	**RNA polyadenylation**	**GO:0032200**	**telomere organization**
GO:0010608	**posttranscriptional regulation of gene expression**	**GO:0034453**	**microtubule anchoring**
GO:0043687	**post-translational protein modification**	**GO:0051054**	**positive regulation of DNA metabolic process**
GO:0032270	**positive regulation of cellular protein metabolic process**	**GO:0006513**	**protein monoubiquitination**
GO:0006417	regulation of translation	**GO:0034660**	**ncRNA metabolic process**
GO:0051247	positive regulation of protein metabolic process	**GO:0051169**	**nuclear transport**
GO:0031401	positive regulation of protein modification process	**GO:0071826**	**ribonucleoprotein complex subunit organization**
GO:0071166	**ribonucleoprotein complex localization**	**GO:0000724**	**double-strand break repair via homologous recombination**
GO:0006473	**protein acetylation**	**GO:0006270**	**DNA replication initiation**
GO:0016573	histone acetylation	GO:0045005	DNA-dependent DNA replication maintenance of fidelity
GO:0006475	internal protein amino acid acetylation	**GO:0006338**	**chromatin remodeling**
GO:0018394	peptidyl-lysine acetylation	**GO:0006275**	**regulation of DNA replication**
GO:0018393	internal peptidyl-lysine acetylation		

Chapter B

Supplementary Table B1. Species in the data set and accession codes of Genome assemblies. When the proteome was retrieved from an online source different from NCBI, the whole accession link is present. Phyla are in alphabetical order, with the four non-Metazoa phyla at the bottom of the table.

Phylum	Species	Genome code / database
ANNELIDA	*Capitella teleta*	GCA_000328365.1
	Dimorphilus gyrociliatus	GCA_904063045.1
	Helobdella robusta	GCF_000326865.1
	Owenia fusiformis	GCA_903813345.1
ARTHROPODA	*Aphis gossypii*	GCF_004010815.1
	Apis mellifera	GCF_003254395.2
	Centruroides sculpturatus	GCF_000671375.1
	Cloeon dipterum	GCA_902829235.1
	Cryptotermes secundus	GCF_002891405.2
	Ctenocephalides felis	GCF_003426905.1
	Daphnia magna	GCF_003990815.1
	Dermatophagoides pteronyssinus	GCF_001901225.1

	Drosophila melanogaster	GCF_000001215.4
	Eurytemora affinis	GCF_000591075.1
	Folsomia candida	GCF_002217175.1
	Hyalella azteca	GCF_000764305.1
	Ixodes scapularis	GCF_002892825.2
	Lepeophtheirus salmonis	GCF_016086655.3
	Limulus polyphemus	GCF_000517525.1
	Nymphon striatum	GCA_016618385.1
	Parasteatoda tepidariorum	GCF_000365465.2
	Penaeus vannamei	GCF_003789085.1
	Sitophilus oryzae	GCF_002938485.1
	Tetranychus urticae	GCF_000239435.1
	Varroa destructor	GCF_002443255.1
BRACHIOPODA	*Lingula anatina*	GCF_001039355.2
BRIOZOA	*Bugula neritina*	GCA_010799875.2
CHORDATA	*Acipenser ruthenus*	GCF_010645085.1
	Amblyraja radiata	GCF_010909765.1
	Branchiostoma floridae	GCF_000003815.2
	Ciona intestinalis	GCF_000224145.3
	Danio rerio	GCF_000002035.6
	Gallus gallus	GCF_000002315.6
	Gekko japonicus	GCF_001447785.1
	Homo sapiens	GCF_000001405.39
	Oikopleura dioica	GCA_000209555.1
	Petromyzon marinus	GCF_010993605.1
	Styela clava	GCF_013122585.1
	Xenopus tropicalis	GCF_000004195.4
CNIDARIA	*Acropora digitifera*	GCF_000222465.1
	Actinia tenebrosa	GCF_009602425.1
	Aurelia aurita	https://marinegenomics.oist.jp/aurelia_aurita/
	Clytia hemisphaerica	http://ftp.ensemblgenomes.org/pub/metazoa/release-52/fasta/clytia_hemisphaerica_gca902728285
	Dendronephthya gigantea	GCF_004324835.1
	Exaiptasia diaphana	GCF_001417965.1
	Henneguya salminicola	GCA_009887335.1
	Hydra vulgaris	GCF_000004095.1
	Morbakka virulenta	https://marinegenomics.oist.jp/morbakka_virulenta/
	Myxobolus squamalis	GCA_010108815.1
	Nematostella vectensis	GCF_000209225.1
	Orbicella faveolata	GCF_002042975.1
	Pocillopora damicornis	GCF_003704095.1
	Stylophora pistillata	GCF_002571385.1
	Thelohanellus kitauei	GCA_000827895.1
CTENOPHORA	*Hormiphora californiensis*	https://github.com/conchoecia/hormiphora
	Mnemiopsis leidyi	https://research.nhgri.nih.gov/mnemiopsis
ECHINODERMATA	*Anneissia japonica*	GCF_011630105.1
	Apostichopus japonicus	GCA_002754855.1
	Asterias rubens	GCF_902459465.1
	Strongylocentrotus purpuratus	GCF_000002235.5

HEMICHORDATA	*Ptychodera flava*	https://marinegenomics.oist.jp/acornworm/
	Saccoglossus kowalevskii	GCF_000003605.2
MOLLUSCA	*Aplysia californica*	GCF_000002075.1
	Biomphalaria glabrata	GCF_000457365.1
	Crassostrea gigas	GCF_902806645.1
	Lottia gigantea	GCF_000327385.1
	Mizuhopecten yessoensis	GCF_002113885.1
	Octopus bimaculoides	GCF_001194135.1
	Pomacea canaliculata	GCF_003073045.1
NEMATODA	*Aphelenchus avenae*	GCA_020875895.1
	Brugia malayi	GCF_000002995.3
	Bursaphelenchus okinawaensis	GCA_904066225.2
	Caenorhabditis elegans	GCF_000002985.6
	Loa loa	GCF_000183805.1
	Necator americanus	GCF_000507365.1
	Plectus sambesii	GCA_002796945.1
	Strongyloides ratti	GCF_001040885.1
	Trichinella spiralis	GCF_000181795.1
NEMERTEA	*Notospermus geniculatus*	https://marinegenomics.oist.jp/nge_v2/
ORTHONECTIDA	*Intoshia linei*	GCA_001642005.1
PHORONIDA	*Phoronis australis*	https://marinegenomics.oist.jp/pau_v2/
PLACOZOA	*Trichoplax adhaerens*	GCF_000150275.1
	Trichoplax spH2	GCA_003344405.1
PLATYHELMINTHES	*Dibothriocephalus latus*	GCA_900617775.1
	Echinococcus granulosus	GCF_000524195.1
	Fasciola hepatica	GCA_002763495.2
	Macrostomum lignano	GCA_002269645.1
	Opisthorchis viverrini	GCF_000715545.1
	Protopolystoma xenopodis	GCA_900617795.1
	Schistosoma mansoni	GCF_000237925.1
	Schmidtea mediterranea	(SMSG.1)
PORIFERA	*Amphimedon queenslandica*	GCF_000090795.1
	Ephydatia muelleri	https://spaces.facsci.ualberta.ca/ephybase/
PRIAPULIDA	*Priapulus caudatus*	GCF_000485595.1
MESOZOA	*Dicyema japonicum*	GCA_011109175.1
ROTIFERA	*Adineta ricciae*	GCA_905250095.1
	Brachionus calyciflorus	GCA_905250105.1
	Didymodactylos carnosus	GCA_905250885.1
	Rotaria socialis	GCA_905332285.1
TARDIGRADA	*Hypsibius dujardini*	GCA_002082055.1
	Ramazzottius varieornatus	GCA_001949185.1
XENACOELOMORPHA	*Praesagittifera naikaiensis*	http://gigadb.org/dataset/100564
	Xenoturbella bocki	https://figshare.com/articles/dataset/Genome_of_Xenoturbella_bocki/
CHOANOFLAGELLATA	*Monosiga brevicollis*	http://ftp.ensemblgenomes.org/pub/protists/release-52/fasta/protists_choanoflagellida1_collection/monosiga_brevicollis_mx1_gca_000002865/
	Salpingoeca rosetta	GCA_000188695.1
FILASTEREA	*Capsaspora owczarzaki*	GCF_000151315.2
ICHTHYOSPOREA	*Chromosphaera perkinsii*	https://figshare.com/articles/dataset/

		Genome_-_Chromosphaera_perkinsii/
	Ichthyophonus hoferi	https://figshare.com/articles/dataset/ Genome_-_Ichthyophonus_hoferi/
	Pirum gemmata	https://figshare.com/articles/dataset/ Genome_-_Pirum_gemmata/
	Sphaeroforma arctica	GCF_001186125.1
PLURIFORMEA	*Corallochytrium limacisporum*	https://figshare.com/articles/dataset/ Genome_-_Corallochytrium_limacisporum/

Supplementary Table B2. Species-specific genomic and genetic statistics used for statistical analyses. Species are ordered in alphabetical order. For calculations, refer to Materials and Methods of Chapter B. **Gene Density** as: Genes over Assembly Length in Mb, i.e. *n* genes per Mb.

Species	Genome Size (Mb)	Genes	Gene Density	Genes / Proteins	G0	G1	G2	*Piwi*-like genes	*Ago*-like genes	Piwi-domain Containing Genes
Acipenser ruthenus	1828.86	36150	197.664	0.538828	40	10	22	4	8	0
Acropora digitifera	449.88	26073	579.555	0.769614	11	5	21	3	5	0
Actinia tenebrosa	234.72	19980	851.227	0.738987	18	3	17	3	2	0
Adineta ricciae	176.04	64538	366.61	NA	35	8	59	9	35	5
Amblyraja radiata	2562.36	18512	722.459	0.475496	19	4	14	3	4	0
Amphimedon queenslandica	166.26	20788	125.033	0.862	8	1	17	2	2	0
Anneissia japonica	586.8	21084	359.305	0.643315	13	3	14	2	3	0
Aphelenchus avenae	264.06	43185	163.542	NA	12	3	16	0	23	46
Aphis gossypii	293.4	12828	437.219	0.693518	13	2	22	6	5	0
Apis mellifera	224.94	9935	441.673	0.423288	12	3	10	2	2	0
Aplysia californica	929.1	19405	208.858	0.727433	12	2	18	2	2	0
Apostichopus japonicus	801.96	30221	376.839	NA	18	4	20	3	2	0
Asterias rubens	420.54	16079	382.342	0.668593	13	3	15	2	1	0
Aurelia aurita	381.42	28604	749.934	0.752598	11	2	20	2	3	0
Biomphalaria glabrata	919.32	25552	277.945	0.696714	11	2	21	2	1	0
Brachionus calyciflorus	117.36	24328	207.294	NA	7	3	22	4	3	0
Branchiostoma floridae	518.34	26689	514.894	0.620083	11	4	18	6	2	0
Brugia malayi	97.8	11371	116.268	0.991196	9	4	17	0	6	1
Bugula neritina	215.16	25318	117.671	NA	14	4	20	2	3	0
Bursaphelenchus okinawaensis	68.46	14593	213.161	NA	5	2	7	1	4	9
Caenorhabditis elegans	97.8	21903	223.957	0.722847	7	2	8	1	9	15
Capitella teleta	332.52	31978	961.687	NA	13	3	13	3	1	0
Capsaspora owczarzaki	29.34	8792	299.659	NA	9	2	1	0	0	0
Centruroides sculpturatus	929.1	24591	264.675	0.692139	15	3	24	4	23	0
Chromosphaera perkinsii	39.12	12463	318.584	NA	3	1	2	1	1	0
Ciona intestinalis	117.36	13713	116.846	0.649936	11	3	9	2	1	0
Cloeon dipterum	176.04	30161	171.33	NA	22	4	39	12	10	4
Clytia hemisphaerica	449.88	19149	425.647	0.739486	13	2	19	2	2	0
Corallochytrium limacisporum	244.5	7535	30.818	NA	4	0	1	0	0	0
Crassostrea gigas	645.48	31371	48.601	0.495272	11	4	19	2	3	0

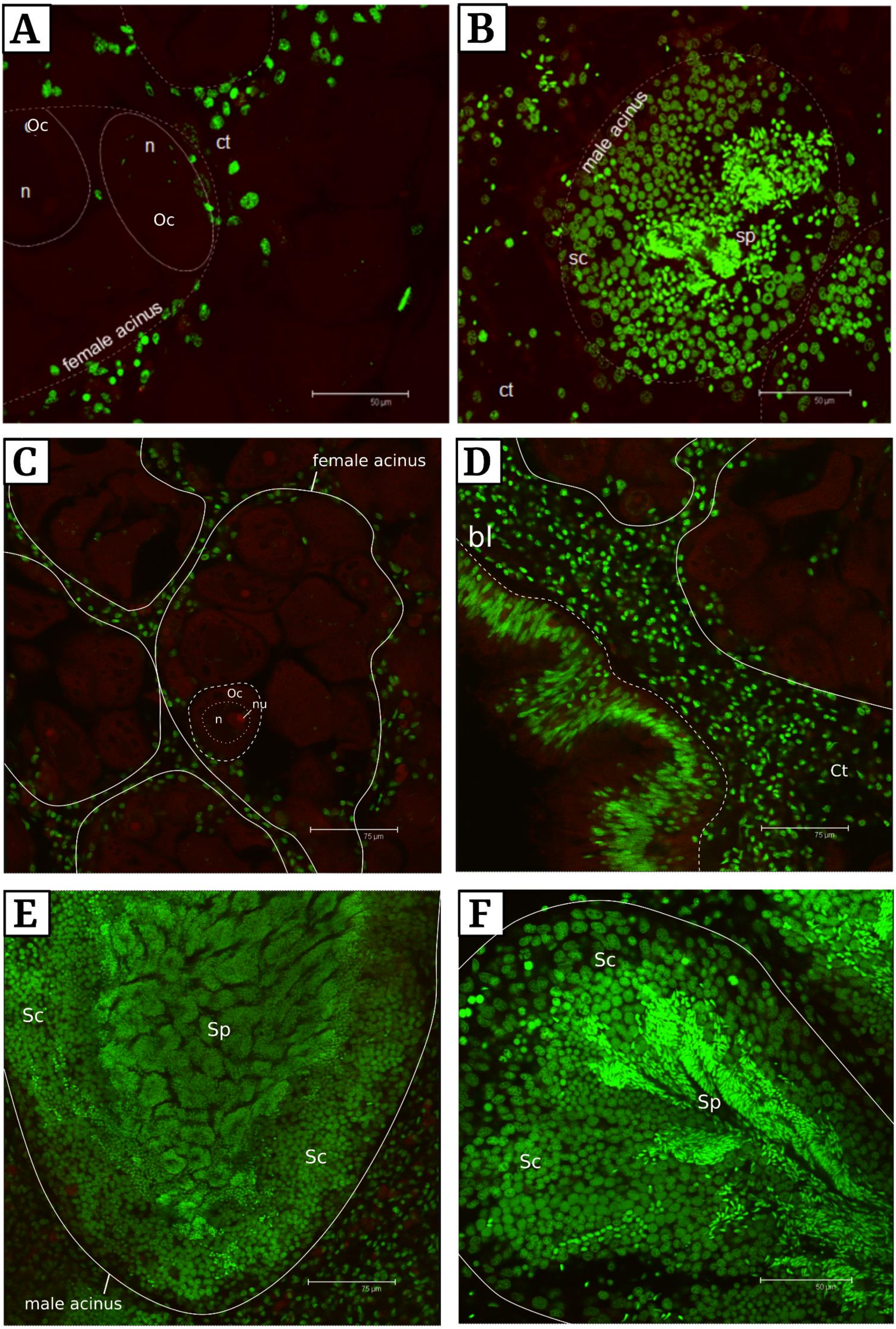
Cryptotermes secundus	1017.12	13170	129.483	0.449718	10	3	12	3	12	0
Ctenocephalides felis	772.62	18878	244.337	0.859889	17	3	21	5	2	0
Danio rerio	1369.2	26533	193.785	0.463961	23	4	15	2	5	0
Daphnia magna	127.14	15351	120.741	0.650935	13	3	19	6	3	0
Dendronephthya gigantea	283.62	22045	777.272	0.767023	12	3	19	1	4	0
Dermatophagoides pteronyssinus	68.46	11184	163.365	0.871232	8	2	10	0	8	10
Dibothriocephalus latus	528.12	19966	378.058	NA	5	2	1	0	5	0
Dicyema japonicum	68.46	4743	692.813	0.946329	0	1	2	4	0	0
Didymodactylos carnosus	371.64	46863	126.098	NA	16	7	30	8	20	2
Dimorphilus gyrociliatus	78.24	16378	209.33	NA	9	4	15	2	1	0
Drosophila melanogaster	146.7	13955	951.261	0.454309	8	4	16	3	2	0
Echinococcus granulosus	107.58	11319	105.215	NA	6	2	2	0	3	0
Ephydatia muelleri	322.74	39329	121.86	NA	8	1	28	2	0	0
Eurytemora affinis	391.2	20716	52.955	0.680887	13	3	27	4	5	0
Exaiptasia diaphana	254.28	22509	885.205	0.811047	16	3	18	4	3	0
Fasciola hepatica	1134.48	11217	988.735	NA	5	2	2	0	4	0
Folsomia candida	224.94	24221	107.678	0.652611	11	3	45	19	11	1
Gallus gallus	1066.02	17576	164.875	0.353074	20	4	11	2	3	0
Gekko japonicus	2493.9	19548	783.833	0.796707	20	4	14	3	4	0
Helobdella robusta	234.72	23426	99.804	NA	11	2	14	2	2	0
Henneguya salminicola	58.68	8187	139.519	NA	1	2	4	2	2	0
Homo sapiens	3100.26	20331	655.784	0.166221	22	5	13	4	4	0
Hormiphora californiensis	97.8	11987	122.566	0.679998	8	2	16	2	3	4
Hyalella azteca	547.68	18608	33.976	0.81797	10	2	11	6	4	0
Hydra vulgaris	850.86	20055	235.703	0.911881	11	1	17	3	2	0
Hypsibius dujardini	107.58	20860	193.902	NA	8	1	16	5	4	0
Ichthyophonus hoferi	88.02	6351	721.541	NA	6	1	2	1	2	0
Intoshia linei	39.12	8724	223.006	NA	4	2	3	1	3	0
Ixodes scapularis	2083.14	24501	117.616	0.75221	13	4	22	4	5	0
Lepeophtheirus salmonis	645.48	14014	21.711	0.678841	11	4	13	4	5	0
Limulus polyphemus	1828.86	22873	125.067	0.591309	19	3	20	2	3	0
Lingula anatina	410.76	27068	658.974	0.653943	16	5	25	5	2	0
Loa loa	88.02	16281	184.969	NA	8	2	17	0	5	3
Lottia gigantea	361.86	23827	658.459	0.999832	10	3	15	2	1	0
Macrostomum lignano	762.84	49018	642.572	NA	22	9	66	14	12	0
Mizuhopecten yessoensis	987.78	24532	248.355	0.59018	13	3	15	2	1	0
Mnemiopsis leidyi	146.7	16548	112.802	NA	13	1	19	4	4	0
Monosiga brevicollis	39.12	9172	234.458	NA	5	2	1	0	0	0
Morbakka virulenta	948.66	24278	255.919	0.837663	11	2	20	2	1	0
Myxobolus squamalis	39.12	5723	146.293	NA	1	2	2	2	2	0
Necator americanus	244.5	19153	783.354	NA	6	2	6	2	9	2
Nematostella vectensis	352.08	23845	677.261	0.694967	20	4	24	5	3	1
Notospermus geniculatus	860.64	43294	503.044	NA	19	6	30	7	4	1
Nymphon striatum	743.28	10384	139.705	0.363955	13	2	8	2	3	0
Octopus bimaculoides	2337.42	15842	677.756	0.660248	11	3	16	3	3	0
Oikopleura dioica	48.9	13527	276.626	NA	3	2	7	1	13	0

Opisthorchis viverrini	616.14	16356	265.459	NA	8	1	2	0	4	0
Orbicella faveolata	489	25929	530.245	0.795685	19	3	20	2	7	0
Owenia fusiformis	498.78	31127	624.063	0.898456	12	3	19	3	1	0
Parasteatoda tepidariorum	1447.44	18601	12.851	0.676007	16	3	14	2	8	0
Penaeus vannamei	1662.6	24987	150.289	0.750969	10	3	12	1	3	0
Petromyzon marinus	1085.58	17580	161.941	0.467852	15	4	14	2	3	0
Phoronis australis	498.78	20473	410.462	NA	12	4	18	3	1	0
Pirum gemmata	88.02	21835	248.069	NA	2	0	1	0	2	0
Plectus sambesii	185.82	40530	218.114	NA	19	3	50	2	12	7
Pocillopora damicornis	234.72	19935	84.931	0.791605	13	3	18	2	2	0
Pomacea canaliculata	440.1	21144	480.436	0.523483	11	3	17	2	1	0
Praesagittifera naikaiensis	655.26	22143	337.927	0.907128	11	1	23	3	4	0
Priapulus caudatus	508.56	15101	296.936	0.729799	10	4	18	2	2	0
Protopolystoma xenopodis	616.14	37906	615.217	NA	3	0	0	0	8	0
Ptychodera flava	997.56	34637	347.217	0.999711	10	1	22	3	1	0
Ramazzottius varieornatus	58.68	19533	332.873	0.849002	6	1	13	5	8	0
Rotaria socialis	156.48	34499	220.469	NA	16	5	33	8	18	3
Saccoglossus kowalevskii	772.62	20935	270.961	0.945915	9	4	17	4	1	0
Salpingoeca rosetta	58.68	11618	197.989	0.990874	7	2	1	0	0	0
Schistosoma mansoni	361.86	10719	29.622	0.914512	8	2	2	0	3	0
Schmidtea mediterranea	704.16	22090	313.707	0.715234	7	1	15	3	3	0
Sitophilus oryzae	772.62	15057	194.882	0.640778	10	3	15	2	2	0
Sphaeroforma arctica	117.36	18213	155.189	0.972397	2	2	1	0	3	0
Strongylocentrotus purpuratus	919.32	27435	298.427	0.713728	14	6	20	3	2	0
Strongyloides ratti	39.12	12445	318.124	NA	2	0	9	0	6	11
Styela clava	342.3	19953	58.291	0.727389	12	2	12	2	1	0
Stylophora pistillata	400.98	24846	619.632	0.747203	13	3	15	2	6	0
Tetranychus urticae	88.02	11686	132.765	0.745091	11	2	21	7	10	0
Thelohanellus kitauei	146.7	15020	102.386	NA	2	0	3	1	2	0
Trichinella spiralis	58.68	16380	279.141	NA	1	2	12	0	47	30
Trichoplax adhaerens	107.58	11518	107.065	0.999826	8	1	5	0	1	0
Trichoplax spH2	97.8	12174	124.479	NA	12	1	5	0	4	0
Varroa destructor	371.64	10260	276.074	0.339499	6	3	12	0	10	1
Xenopus tropicalis	1447.44	21898	151.288	0.48478	22	5	14	4	4	0
Xenoturbella bocki	537.9	24134	448.671	NA	6	1	23	5	4	0

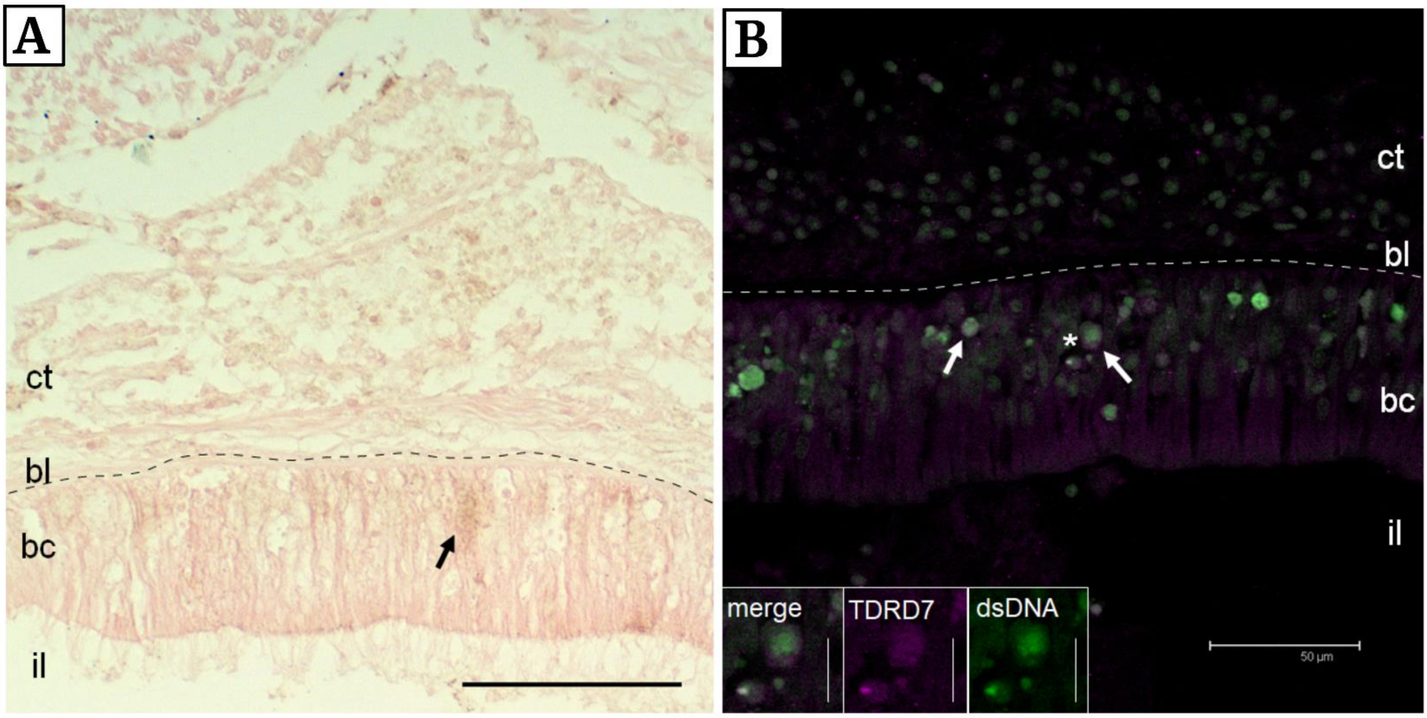
Chapter C

Supplementary Table C1. PCR primers. All the 13 primer couples used are reported in the table. For the first and last sequence portion, we developed three couples since the *in silico* prediction had bad parameters and we wanted to be sure to amplify the portion: all six of them worked. The PCR cycle was: 94°C for 30s, 55°C for 30s; 73°C for 90s. This was repeated for a minimum of 30 and a maximum of 37 times, depending on the product yield. The reaction included 2 minutes at 94°C at the beginning and 5 minutes at 73°C at the end. The only difference between the reactions for the primer couples was the annealing temperature (see last column).

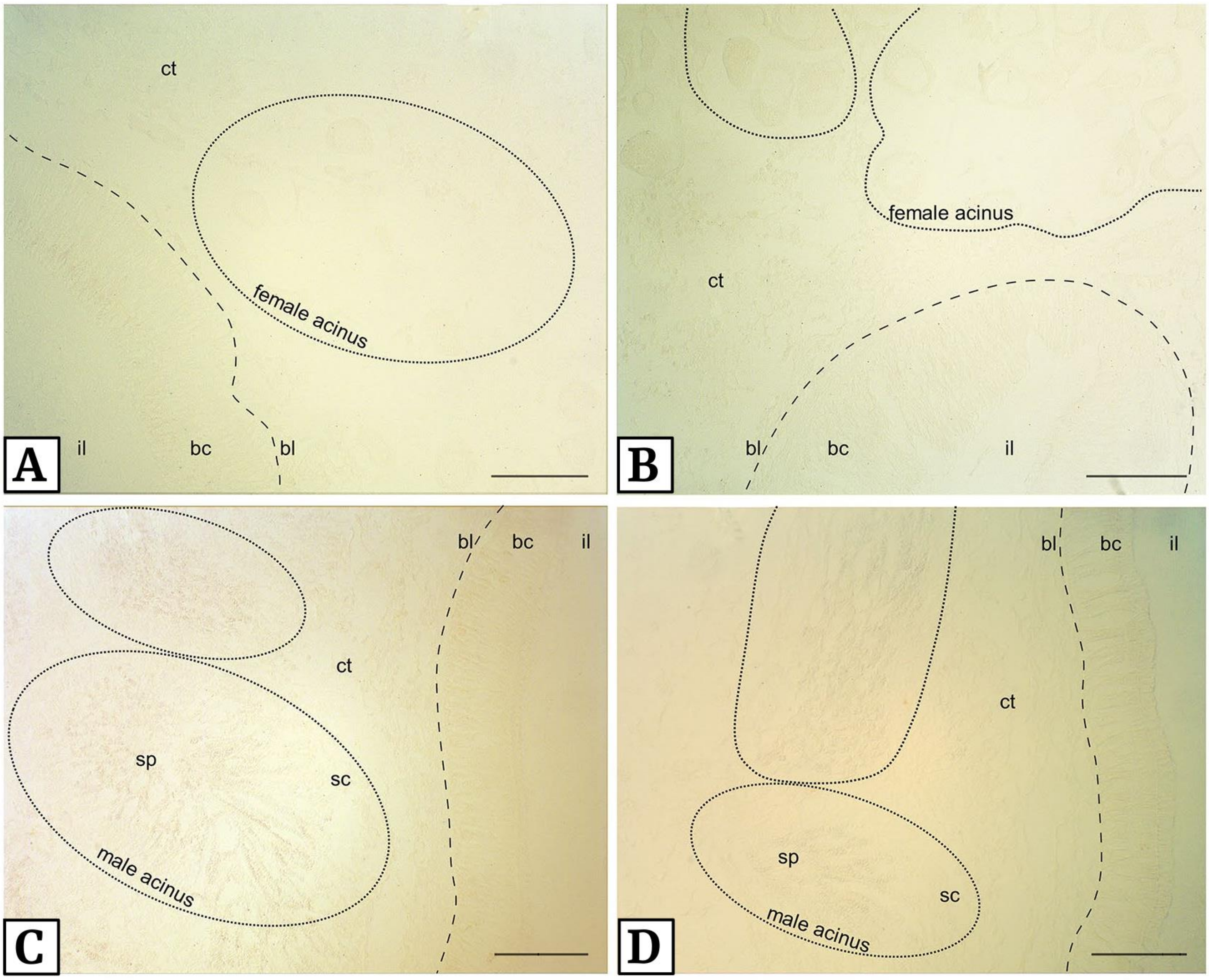
Name	Primer	Sequence	Length	Start	Stop	Annealing T (°C)
1a	Forward	GTCGTTCGAAAAGAGGCTGC	20	73	92	55
1a	Reverse	CTGTACTGACTTGAGCCCCG	20	631	612	55
1b	Forward	CGAAAAGAGGCTGCATTCGT	20	79	98	55
1b	Reverse	GGGGTAGAAATCGTGACCCG	20	609	590	55
1c	Forward	CGAAGTCGTTCGAAAAGAGGC	21	69	89	55
1c	Reverse	CCGTGGGGGTAGAAATCGTG	20	614	595	55
2	Forward	TCCACTTCACACATCCAGGC	20	507	526	55
2	Reverse	TCCCTACCGTCTCCTCCTTG	20	1242	1223	55
3	Forward	GCCATGTTGGAAGGTGGAGA	20	957	976	58
3	Reverse	CTCGGTGGACTGCTTGTTGA	20	1651	1632	58
4	Forward	AAGTACAATGAGGACCCGCC	20	1353	1372	58
4	Reverse	ATGGCCTGGCACGGATATTT	20	1849	1830	58
5	Forward	TCAACAAGCAGTCCACCGAG	20	1632	1651	55
5	Reverse	TGCTCTGGAATCAGCTCGTC	20	2401	2382	55
6	Forward	CCGTGATGCTCAGACTGGTT	20	2164	2183	55
6	Reverse	AATTTGCCGGACTCGGTGAT	20	2700	2681	55
7	Forward	GTTGCCGAGGTTGTTGACAG	20	2541	2560	55
7	Reverse	CCCGGACTTCGACCCAAAAT	20	3004	2985	55
8	Forward	ACACCGAGAACCGAATCACC	20	2667	2686	55
8	Reverse	CTGCCAAGTGTTGCGTTGAA	20	3418	3399	55
9a	Forward	GTCTGTTCCGAGACCATTCCA	21	3032	3052	55
9a	Reverse	TCTCGAAAGGAGTCTTTAGCAC	22	3701	3680	55
9b	Forward	CCGAGACCATTCCAGATTCCA	21	3039	3059	55
9b	Reverse	TTCTCGAAAGGAGTCTTTAGCAC	23	3702	3680	55
9c	Forward	CTGTTCCGAGACCATTCCAGA	21	3034	3054	55
9c	Reverse	CTCGAAAGGAGTCTTTAGCACTT	23	3700	3678	55


A
Oc
n
n
ct
Oc
female acinus
50 µm
B
male acinus
sp
sc
ct
50 µm
C
female acinus
Oc
nu
n
75 µm
D
bl
Ct
75 µm
E
Sc
Sp
Sc
male acinus
75 µm
F
Sc
Sp
Sc
50 µm

Supplementary Figure C1. IF control samples with omission of primary antibodies. (Figure in previous page) Both anti-chicken Dylight®550 (for clam-specific antibodies; **A-B**) and anti-rabbit AlexaFlour®48 (for human-built antibodies; **C-D-E-F**) controls are depicted in the figure. No staining is present in any of the sections. Female and male acini histology is visible, together with the close gonad association to the intestine. Bl = basal lamina; ct = connective tissue; Oc = oocyte; n = nucleus; nu = nucleolus; Sc = spermatocytes; Sp = spermatozoa. Green: TO-PRO3 dsDNA dye.



Supplementary Figure C2. Immunolocalisation of clam-specific antibodies during the spent phase. Given the lack of gametes, sexes were not distinguishable. **A**: IHC localisation; a very light staining is present in few intestinal cells (arrow); scale bar = 100 µm. **B**: IF localisation; also in this case, very few cells are very lightly stained (arrows; asterisk tag the section highlighted in the bottom inset. Purple: Anti-TDRD7 staining; Green: TO-PRO3 dsDNA dye. Bc = batiprismatic cells; bl = basal lamina; ct = connective tissue; il = intestinal lumen.



Supplementary Figure C3. IHC control samples with omission of primary antibodies. Both HRP anti-chicken (for clam-specific antibodies; **A-C**) and HRP anti-rabbit (for human-built antibodies; **B-D**) controls are depicted in the figure. No staining is present in any of the sections. bl = basal lamina; ct = connective tissue; il = intestinal lumen; Sc = spermatocytes; Sp = spermatozoa.

Appendix

In my first year of my PhD, the work on the projects presented in this thesis overlapped with the work started at the University of Bologna during my Master Thesis. The taxa, tools, and methods used were coherent with some of those presented in the thesis chapter, allowing me to strengthen the bioinformatic skills that were useful for the subsequent PhD program. This work regarded mitonuclear coevolution, that was investigated in Bivalvia, a Mollusca class with some interesting features regarding mtDNA. The results were later published (https://doi.org/10.1093/molbev/msab054) and the subsequent appendix represents a reduced version of the paper. For Supplementary Material, refer to the online version of the manuscript.

Mitonuclear Coevolution, but not Nuclear Compensation, Drives Evolution of OXPHOS Complexes in Bivalves

Introduction

Mitochondria are the product of an ancient endosymbiotic event between an Archaea-like prokaryote and an alpha-proteobacterium (reviewed in Martin et al. 2015) that led to the evolution of eukaryotes and morphologically complex life as we know it today (Martin and Müller 1998; Martin and Koonin 2006; Lane and Martin 2010; Hill 2015; Zachar and Szathmáry 2017). The mitochondrial genome is a genetic relic of complex evolutionary processes that resulted in an extensive reduction of the alpha-proteobacterium genome, involving both gene loss and transfer to the nuclear genome (Gray et al. 1999; Timmis et al. 2004; Martin and Koonin 2006; Gray 2012).

At present, different eukaryotic lineages have variable mitochondrial genome sizes, organization, and gene content (Kolesnikov and Gerasimov 2012; Sloan et al. 2018). However, beside genes involved in translation, one consistent pattern is the maintenance of a limited set of Protein Coding Genes (PCGs) involved in the Oxidative Phosphorylation (OXPHOS) metabolic pathway, the main mechanism of ATP production in aerobic eukaryotes. OXPHOS is carried out by four protein complexes that produce a proton gradient across the internal mitochondrial membrane (Complexes I to IV, or CI-IV), and an ATPase that exploits this gradient to produce ATP (Complex V, or CV). In almost all bilaterian animals, 13 PCGs encoding components of CI and CIII-CV are found in the mitochondrial genome. In Metazoa the number of nuclear-encoded subunits is variable but ranges around 70, with Complex II being composed entirely of nuclear-encoded proteins.

One of the consequences of this binary genome delegation for such a critical metabolic process is that mitochondrial and nuclear genomes products must physically interact for proper OXPHOS functioning. However, these two genomes experience different evolutionary dynamics: for instance, mitochondria have a small effective population size, are uniparentally inherited, and often experience higher substitution rates (see Ballard and Whitlock 2004). This has led to a general prediction of mitonuclear coevolution: evolution in one genome should select for complementary changes in the other to ensure correct mitochondrial functions (Rand et al. 2004; Bar-Yaacov et al. 2012; Hill 2019; Hill 2020). Probably the most persuasive evidence of the tight coevolution of mitochondrial and nuclear OXPHOS genes comes from experiments with cytoplasmic hybrids. In these experiments, divergent mitochondrial genomes are expressed against foreign nuclear backgrounds via experimental crossing designs or nuclear transfer, often causing OXPHOS inefficiency and lowered fitness (McKenzie et al. 2003; Niehuis et al. 2008; Burton and Barreto 2012; Barreto and Burton 2013; Barreto et al. 2018; Healy and Burton 2020).

Signatures of mitonuclear coevolution are also present in the molecular evolution of OXPHOS genes. In insects rates of evolution are strongly correlated in mitochondria-encoded and nuclear-encoded OXPHOS (mtOXPHOS and nuOXPHOS, respectively) genes, but not between mitochondrial genes and nuclear-encoded genes lacking mitochondrial interactions (Yan et al. 2019). Such Evolutionary Rate Correlation (ERC) in mitochondrial genes and their nuclear-encoded counterparts generally extends across eukaryotes: lineages with fast-evolving mitochondrial genes also have fast-evolving mitochondria-interacting nuclear genes (Havird and Sloan 2016).

However, why mitonuclear coevolution is common and whether it is adaptive are less thoroughly understood. Some have argued that increased dN/dS ratios (i.e. ratio between nonsynonymous substitutions per nonsynonymous site and synonymous substitutions per synonymous site, also known as ω) in nuOXPHOS genes of animals are due to relaxed functional constraints on peripheral nuOXPHOS subunits, not positive selection in response to mitochondrial changes (Nabholz et al. 2013; Popadin et al. 2013; Zhang and Broughton 2013). Closely related taxa in the angiosperm genus *Silene* with highly divergent mitochondrial mutation rates have proven valuable in addressing these hypotheses. In taxa with fast mitochondrial mutation rates, nuOXPHOS subunits show elevated dN/dS ratios as a result of positive selection, despite still acting as peripheral subunits (Sloan et al. 2014; Havird et al. 2015; Havird et al. 2017). Structural information has also been used to show that nuOXPHOS changes tend to occur in areas that interact with mitochondria-encoded residues (Osada and Akashi 2012; Havird et al. 2015). These results are consistent with the most popular hypothesis stemming from mitonuclear coevolution: nuclear compensation, which posits that inefficient selection in mitochondrial genomes causes mildly deleterious mutations to accumulate, which are offset by compensatory changes in interacting nuclear-encoded genes. However, direct evidence for nuclear compensation over other forms of mitonuclear coevolution remains scarce, especially in invertebrates.

Here, we examine mitonuclear coevolution in Bivalvia, a class of sedentary molluscs. These animals represent an interesting observational unit for such studies for several reasons. First, bivalve phylogenies inferred with mitochondrial DNA (mtDNA) show discordance with nuclear ones, mainly with regards to deep relationships between Pteriomorphia, Palaeoheterodonta and Heterodonta (Doucet-Beaupré et al. 2010; Bieler et al. 2014; González et al. 2015; Plazzi et al. 2016). However, phylogenies based on nuOXPHOS subunits are lacking and phylogenetic concordance in these specific nuclear-encoded genes could be a consequence of mitonuclear coevolution. Moreover, bivalves include a unique and evolutionarily stable exception to the strictly maternal inheritance (SMI) of mitochondria in animals: more than 100 species (Gusman et al. 2016) show doubly uniparental inheritance (DUI), with a maternally-transmitted mtDNA (F-type) and a paternally-transmitted mtDNA (M-type) (see Zouros 2013 for a review). The amino acid p-distance between F- and M-type mtOXPHOS proteins can be >50% (Doucet-Beaupré et al. 2010; Zouros 2013) and both F- and M-type mtDNA and their products (RNAs and proteins) are present in females and males (i.e., heteroplasmy; see Ghiselli et al. 2019 for a thorough discussion). Such peculiar mitochondrial inheritance implies that the same nuclear background has to co-function with two different mitochondrial genomes, adding another layer of complexity to mitonuclear coevolution. Bivalves also show variation in rates of mitochondrial evolution, but their sedentary nature suggests maintaining highly efficient OXPHOS may be under weaker selection compared to taxa with higher metabolic requests. Moreover, it appears that bivalve mitochondrial mutation rates are not dramatically higher than the nuclear ones (see for instance Allio et al. 2017), therefore potentially representing a different mitonuclear coevolutionary landscape respect to deeply investigated taxa like vertebrates (where mitochondrial mutation rates can be ~30 times as high as the nuclear ones). Coherently, a recent study by Iannello et al. (2019) observed that mt and nuOXPHOS subunits did not show significantly different dN/dS ratios in two congeneric species of bivalves, one of which has DUI.

To explore mitonuclear coevolution in bivalves, we investigated phylogenetic signals of mt and nuOXPHOS proteins and dN/dS ratios in the OXPHOS complexes spanning the whole phylogenetic tree of Bivalvia, including both SMI and DUI species. We also examined ERCs among mt and nuOXPHOS subunits, as well as nuclear-encoded genes with no mitochondrial interactions as a negative control. Furthermore, we investigated signals of site-specific positive selection in the context of protein structures, mitonuclear interactions, and functional sites.

Results

Dataset and Annotation

Out of the 40 bivalve transcriptomes we selected and assembled (based on proportional and wide phylogenetic sampling; Supplementary Table 1), 9 were excluded either because of low quality of the data or because of massive contamination (losing the only Archiheterodonta available. Out of the 7 DUI species included in the present study, we obtained M-type mtOXPHOS subunits for 4 of them, namely *Cristaria plicata, Hyriopsis cumingii, Mytilus edulis,* and *Ruditapes philippinarum*.

Out of 403 expected mtOXPHOS sequences (13 subunits for 31 species), we could retrieve 343. As regards nuOXPHOS, we could annotate 66 subunits in our data set, and roughly 27.2% of the total expected sequences (66 subunits for 31 species = 2,046 sequences) were missing and the implementation of iterative intra-dataset runs with the PSIBLAST tool only moderately improved recovery of nuOXPHOS subunits (Figure 1). However, the presence/absence patterns of nuOXPHOS subunits were not random in regard to the position of the nuclear-encoded subunits within the complexes. Subunits predicted to contact mtOXPHOS subunits tended to have lower annotation rates than "non-contact" subunits. To summarize, the protein sequence evolution analyses were conducted on 31 bivalve species, for a total of 1,864 sequences (379 mitochondrial and 1,485 nuclear).

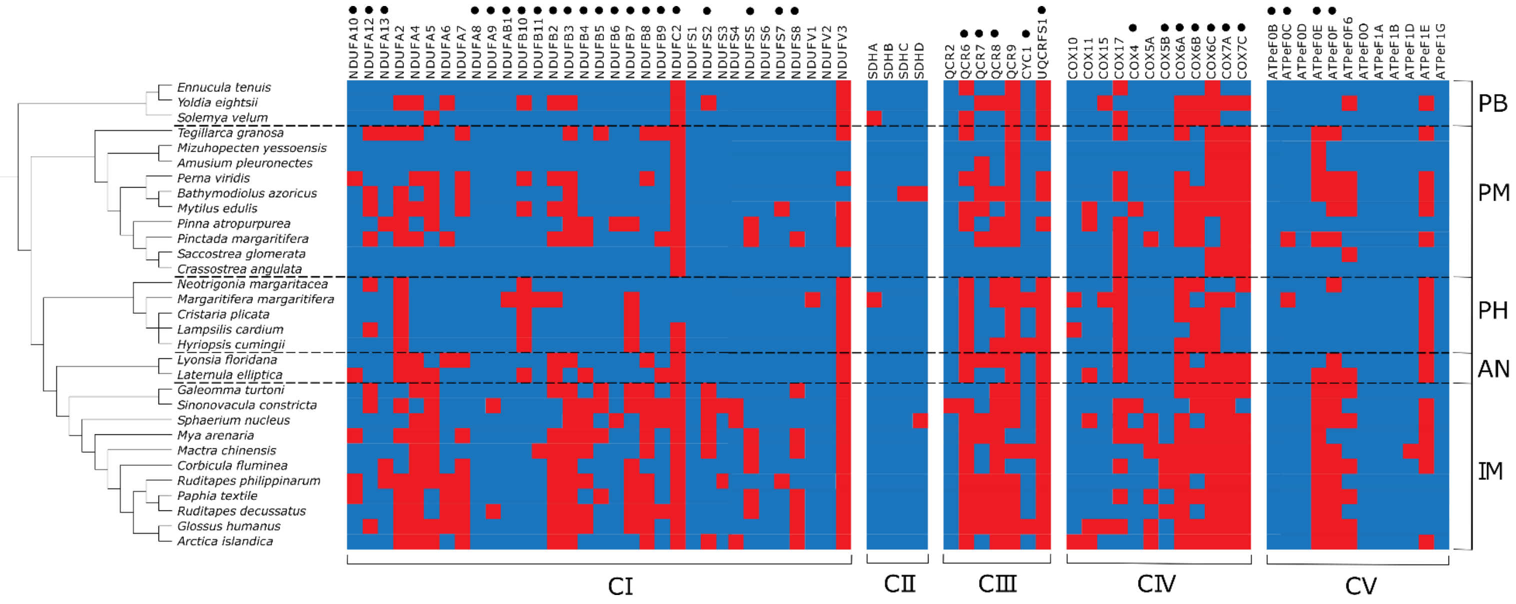


Figure 1: Annotation of nuDNA-encoded OXPHOS subunits. Presence and absence of each subunit in each species are depicted in red and blue, respectively. Left: species tree as built recovering data from literature (see "Evolutionary Rate Correlations" subsection of Materials and Methods for details). Top: protein nomenclature; black dots indicate subunits in contact with mitochondria-encoded proteins. Right: taxonomic clades (PB: Protobranchia; PM: Pteriomorphia; PH: Palaeoheterodonta; AN: Anomalodesmata; IM: Imparidentia). Bottom: respective OXPHOS complex.

Concordance Between mt and nuOXPHOS Phylogenetic Inferences

PartitionFinderProtein estimated LG+G as the best fitting model for all partitions (LG+G+F for mitochondrial partitions). Maximum Likelihood trees were inferred for both the mtOXPHOS and nuOXPHOS concatenated datasets (see Materials and Methods for details).

Paleoheterodonta clustered separately from all other Autobranchia in both datasets. This pattern is common for mitochondrial phylogenies of bivalves, that show topologies in which Euheterodonta (Imparidentia and Anomalodesmata) clusters together with Pteriomorphia (Doucet-Beaupré et al. 2010; Plazzi et al. 2016). However, such relationships have always been a matter of debate, since no phylogenetic analyses based on nuclear markers or genome-wide data had obtained that topology so far, but rather displayed Euheterodonta clustering with Palaeoheterodonta (Kocot et al. 2011; Sharma et al. 2012; Stöger and Schrödl 2013; González et al. 2015). On the other hand, inner relationships among Paleoheterodonta, Imparidentia and Pteriomorphia were mainly concordant between the two trees and with the literature, with some minor exception (e.g. *Yoldia eightsii* position).

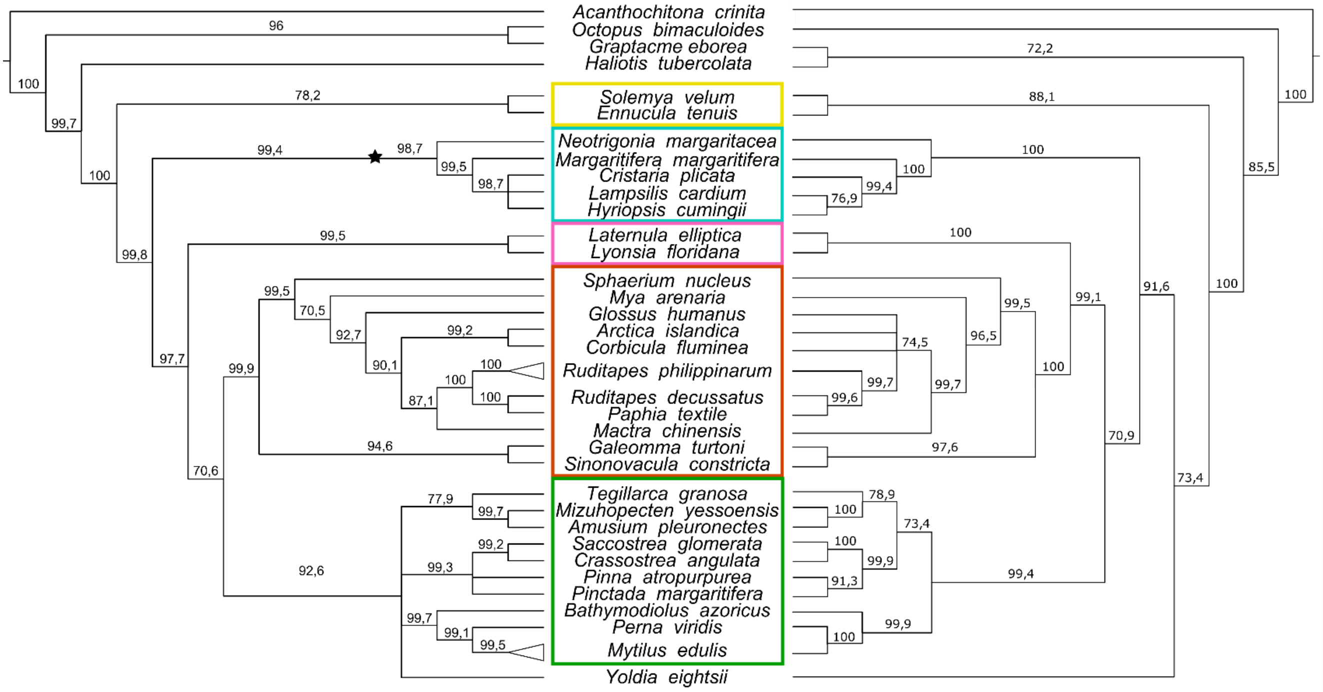


Figure 2: Maximum-likelihood tree inference of mitochondrial and nuclear datasets. Trees were inferred with RAxML v8.2.11 on the two concatenated datasets. Only topologies are depicted in the figure. Bootstrap supports are depicted over each branch (supports lower than 70 were collapsed; 1000 bootstrap replicates were performed). Left: mitochondrial topology (the star represents the omitted branching of unionids male mitochondria-encoded subunits. Other DUI species with both genomes available diverged terminally and the splits were collapsed in triangles, i.e. *R. philippinarum* and *M. edulis*). Right: nuclear topology. Clade acronyms as in Fig. 1.

Strong Correlation Between Evolutionary Rates of mt and nuOXPHOS Proteins

In order to examine coevolutionary signals in mitochondrial and nuclear genes, we performed an Evolutionary Rates Correlation (ERC) analysis. We obtained a “species tree” from the literature (see Materials and Method for details) and optimized branch lengths on that topology for the mtOXPHOS dataset, the nuOXPHOS one, and a third dataset of randomly chosen nuclear orthologues that share no roles in OXPHOS assembly or functioning (all values used for ERC analyses are in Supplementary Material). We then investigated the correlations between the branch lengths (root-to-tip, representing amino acid substitutions) of these three subsets of proteins.

There was a much stronger correlation between the branch lengths of mt and nuOXPHOS subunits compared with the other ERCs (Figure 3A). In particular, nuOXPHOS branch lengths had an almost perfect positive linear correlation of 0.967 with mtOXPHOS branch lengths (95% confidence interval: 0.931-0.984; p < 2.2e−16), while the random nuclear orthologues were only mildly correlated with the mtOXPHOS subunits (ρ = 0.437; 95% confidence interval: 0.098-0.686), although still with statistical significance (p = 1.39e−2, Figure 3B). Moreover, the correlation between nuOXPHOS and random nuclear orthologues was also statistically significant, but much lower than the correlation between nuOXPHOS and mtOXPHOS subunits (ρ = 0.548; 95% confidence interval: 0.240-0.756, p = 1.42e−3; Supplementary Figure 1).

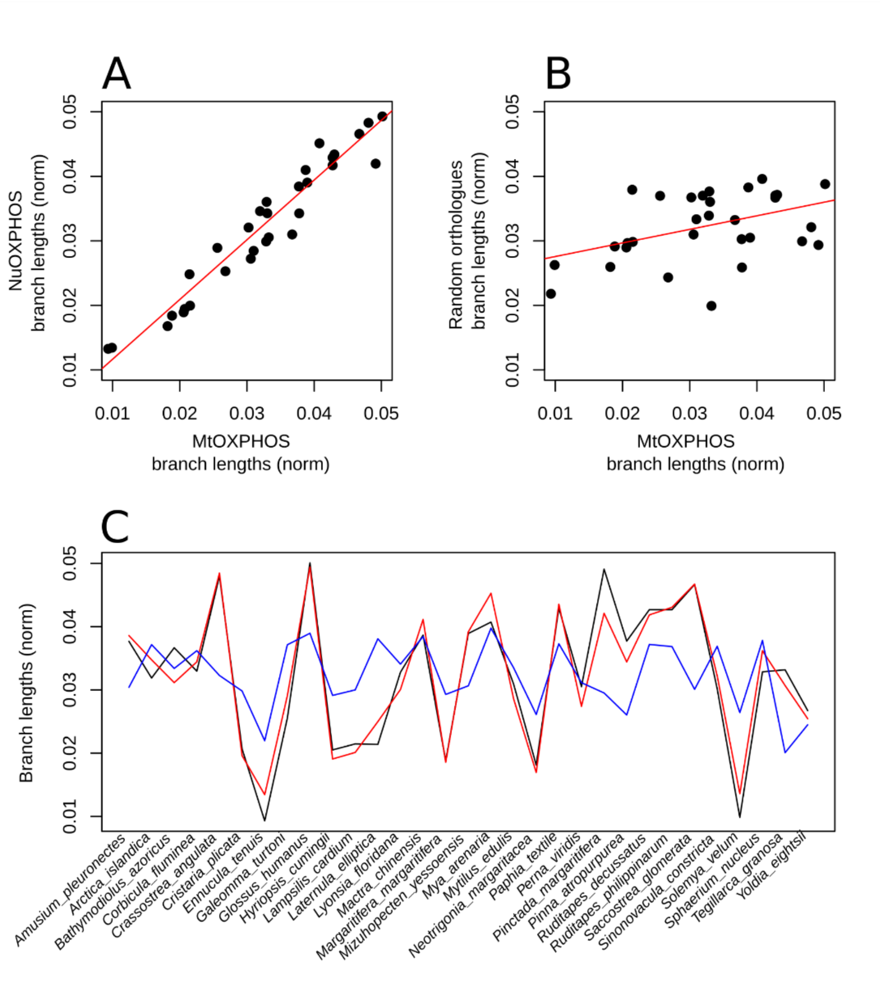


Figure 3: Evolutionary rate correlations analysis. a-b: correlation graphs between normalized branch lengths (per cumulative length of each tree, i.e. cumulative sum of branch lengths for each tree = 1) of mtOXPHOS subunits vs nuOXPHOS subunits (ρ = 0.967; 95% confidence interval: 0.931-0.984; p = 2.2e−16), and mtOXPHOS subunits vs random orthologues (ρ = 0.437; 95% confidence interval: 0.098-0.686 p = 1.39e−2), respectively. **c**: the black line represents the values of normalized mtOXPHOS branch lengths for each species in both graphs, the red line follows the values of normalized branch lengths on the same species for nuOXPHOS (average difference = 0.00219), and the blue line represents the branch lengths of random orthologues (average difference = 0.00819). This graph is useful to visualize the greater average difference in random orthologues' branch lengths with respect to mtOXPHOS ones, compared to the differences between the latter and nuOXPHOS subunits. The lines that link the species are virtual and their purpose is simply to highlight the differences in the three relative trends of branch lengths.

The uniformity in the "nuclear signal" represented by the random orthologues was checked by dividing them in two random subsets (1,000 random divisions were performed) and assessing that there is a strong correlation between them (median ρ value for the 1,000 random subsets = 0.886 median p = 3.36e−11). Moreover, the strong correlation between mtOXPHOS and nuOXPHOS subunits held after normalizing the two distributions for each subset of random orthologues as an attempt to control for variation in overall rates of nuclear genome evolution among species (median ρ value for the 1,000 iterations = 0.925; median p = 9.99e-14; Supplementary Figure 2). The mitonuclear OXPHOS correlation was also robust after calculating it only using the terminal branches of each species (therefore avoiding any possible within-distribution bias): ρ = 0.937, p = 9.43e−15.

After normalizing the branch lengths of each tree for the total length of the trees themselves, it became clear that the curve trend of the nuOXPHOS proteins was more similar to that of the mtOXPHOS ones than to that of random orthologues (Figure 3C). For each species, the difference between the normalized branch lengths of mtOXPHOS proteins and nuOXPHOS proteins was on average 3.7 times lower than the difference between mtOXPHOS and random nuclear orthologues.

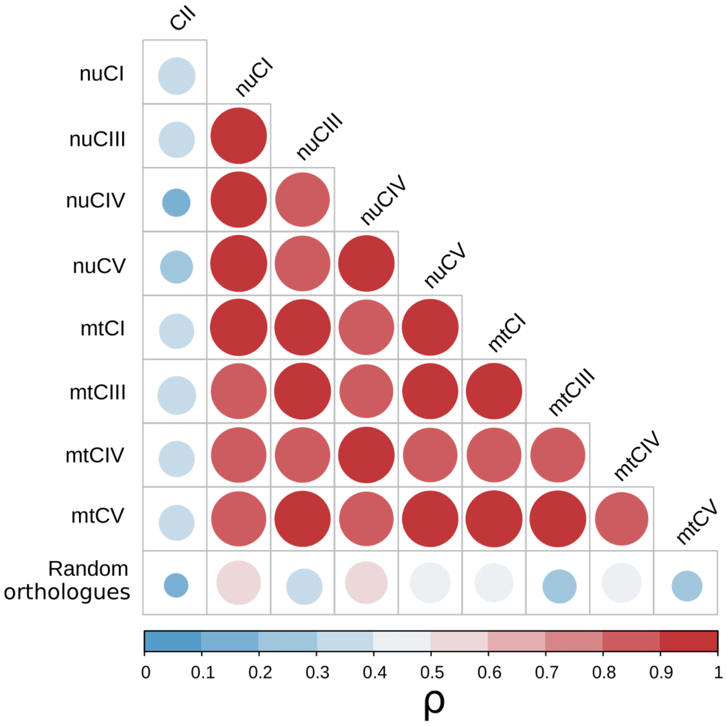


Figure 4: Evolutionary rate correlations between each complex component. Graphical correlation matrix (Pearson's r) between each component of each complex and the random orthologues. CII and the random orthologues dataset shared lower correlation values with all other complex components, which were generally all consistently correlated with each other.

Another interesting coevolutionary signal resulted from correlation analyses performed for each component of each complex (i.e. when datasets of mt and nuOXPHOS subunits within each complex were correlated, Figure 4). All components, whether mitochondria- or nuclear-encoded, of all complexes were highly correlated with each other (ρ ranging from 0.819 to 0.950), with the exception of CII, that correlated to a very low extent with all other components (the highest ρ is 0.376). Moreover, all correlations with the branch lengths of random orthologues shared low values, and this was true also for CII subunits. With regard to CI, CIII, and CIV, the within-complex mitochondria- and nuclear-encoded components represented reciprocal best correlations. Nevertheless, these ρ values differed for at most 0.02.

Comparable dN/dS in mt and nuOXPHOS Subunits

The distributions of dN/dS varied widely across OXPHOS gene products (Supplementary Figure 4). Interestingly, we observed significantly higher dN/dS values in "contact" nuOXPHOS subunits (median = 0.25) compared to "non-contact" ones (median = 0.1665; Wilcoxon-Mann-Whitney test: $p < 2.2e{-}16$; Figure 5A). Moreover, mtOXPHOS subunits had values of dN/dS almost one order of magnitude higher than what was previously observed in most Metazoa (see for example: Nabholz et al. 2013; Havird and Sloan 2016), with a median value of 0.2241 (Figure 5A). The dN/dS values of mtOXPHOS subunits were significantly higher than those of "non-contact" nuOXPHOS subunits (Wilcoxon-Mann-Whitney test: $p < 2.2e{-}16$), but similar to those of "contact" nuOXPHOS ones (Wilcoxon-Mann-Whitney test: $p = 0.183$; Figure 5A).

Variable evolutionary dynamics were observable when considering each OXPHOS complex separately (summary of the statistical relationships between all complexes distributions is in Supplementary Table 3). Overall, CV displayed the fastest evolution among all complexes, both regarding its mitochondria-encoded components (0.2692 median dN/dS, significantly higher than all other mtOXPHOS subunits) and its nuclear-encoded ones (0.3836 median of "contact" subunits, significantly higher than all other nuOXPHOS subunits, except for nuCIV; Figure 5B). On the other hand, the slowest evolving complex was by far CII, that had lower dN/dS values compared with all other nuclear-encoded components (Figure 5B).

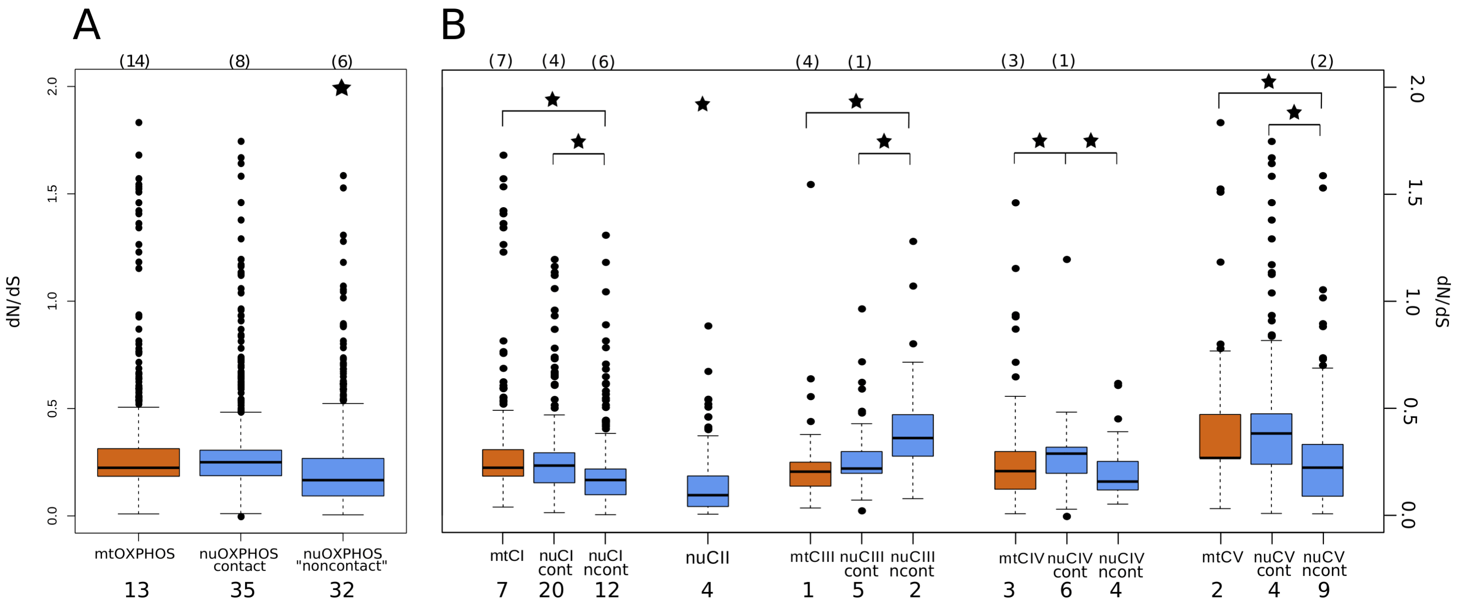


Figure 5: dN/dS distributions of mtOXPHOS subunits, "contact" nuOXPHOS subunits, and "non-contact" nuOXPHOS. MtOXPHOS distributions are depicted in orange, nuOXPHOS ones in blue. Black lines within the boxes are the medians; the two hinges of the boxes approximate the first and the third quartile; whiskers extend to a roughly 95% confidence interval. Outliers are represented as black dots. Stars represent statistical significance of the relationship highlighted (single stars indicate statistically significant differences with all other distributions). Top: number of outliers not depicted in the figure. Bottom: Number of subunits included in the distributions. **A:** Overall distributions. "Non-contact" nuOXPHOS subunits had a distribution statistically lower than both the other distributions (indicated by the star). MtOXPHOS and "contact" nuOXPHOS subunits shared statistically similar distributions. **B:** Distributions of dN/dS for each complex and each compartment. "Contact" subunits are displayed as "cont" while "non-contact" subunits are displayed as "ncont".

Within each complex, “contact” nuOXPHOS subunits showed dN/dS values on average higher than “non-contact” components (with an exception represented by the high dN/dS of “non-contact” nuCIII subunits; Figure 5B). However, rates of evolution of mtOXPHOS and “contact” nuOXPHOS components within a complex were never significantly different (reflecting the relationships observed for the overall datasets, see Figure 5A), in stark contrast to other animals where dN/dS of nuOXPHOS subunits were higher. The only case where “contact” nuOXPHOS subunits had significantly higher values of dN/dS with respect to their mitochondria-encoded counterparts was in CIV.

No Clear Site-Specific Signature of Nuclear Compensation

For all subunits, the LRTs evaluated M3 as a better model compared to M0, meaning that a uniform rate of protein evolution across all sites would not represent the dataset as well as variable rates. We then tested the likelihoods of models that implement distributions of sites under positive selection. For 33 subunits both M2a and M8 resulted as better models (for model descriptions see Materials and Methods; Yang 2007). All these subunits were also analysed for site-specific dN/dS under the MEC model, that includes empirical weights of the different amino acid substitutions detecting no site under putative positive selection. However, comparing corrected Akaike Information Criteria scores, M8 model resulted nevertheless better 9 of the 33 subunits. These 9 subunits with site-specific signatures of positive selection included COX1 alongside nuclear-encoded subunits of Complexes I (NDUFA2, NDUFB2, NDUFS2, and NDUFV1), II (SDHA and SDHB), and V (ATPeF1A and ATPeF1B) (summary of LRTs in Supplementary Table 2; see Materials and Methods for details on this site-specific analysis and Supplementary Figure 6 for a graphic summary).

However, when comparing the site-specific results of positive selection with the annotated sites of catalysis, substrate binding, or subunits interface, there was no correlation. The only exceptions of were observed in COX1, and they represented interaction sites with other subunits: codons 357 (interaction with COX2) and 524 (interaction with COX5B) of the *C. angulata* sequence. Another interesting region consisted in positions 406-410 of NDUFS2, that were positively selected and close to the start of the C-t tail that makes contact with the mitochondria-encoded subunits of the complex. However, these did not represent actual contact sites. Overall, positively selected positions were mostly associated to residues on the surface of complexes, and never buried ones (results of structural alignments with references in Supplementary Table 4; predicted 3D structures with putative positively selected sites in Supplementary Material).

Discussion

Strong Signals of Coevolution Between mt and nuOXPHOS Genes by Phylogenetic inference and Evolutionary Rate Correlations

One of the most contradictory findings in opposition to mitonuclear coevolution is the prevalence of mitonuclear phylogenetic discordance in animals - mitochondrial genes yield one topology while nuclear genes produce another (Sharma et al. 2012; Toews and Brelsford 2012). Bivalves are no exception, and the major difference between the previous phylogenies inferred with the mitochondrial vs. nuclear markers lies in the deep relationships. This may be due to incomplete lineage sorting, mitochondrial introgression, or errors in reconstructing phylogenies. One resolution to this contradiction may be that nuclear phylogenies are often based on anonymous loci (e.g., SNPs obtained by RAD-Seq) or genes that lack mitochondrial interactions. Accordingly, it has been suggested that nuOXPHOS genes should show more similar phylogenetic signals to mtOXPHOS genes under mitonuclear coevolution compared with non-interacting nuclear genes (Sloan et al. 2017). The results of our phylogenetic analysis (Figure 2) were consistent with such predictions, since our nuOXPHOS phylogeny was more similar to the mtOXPHOS phylogeny than previous topologies based on either a handful of nuclear markers or transcriptome-wide analyses (Bieler et al. 2014; González et al. 2015).

Another strong signal of mitonuclear coevolution was the almost perfect positive linear correlation between branch lengths calculated on the same species tree for mtOXPHOS and nuOXPHOS subunits ($\rho = 0.967$; Figure 3A). Random nuclear orthologues lacking mitochondrial interactions were only mildly correlated to mtOXPHOS subunits ($\rho = 0.437$; Figure 3B), suggesting that genome-wide changes in evolutionary rates only partially explain the strong ERC between mt and nuOXPHOS genes. Similar results were previously found for insects (e.g.: Yan et al. 2019) and between plastid-encoded and plastid interacting genes in angiosperms (Williams et al. 2019). Such strong ERCs between mitochondrial and mitochondria-interacting nuclear genes represent some of the strongest evidence of shared evolutionary dynamics between the mitochondrial and the nuclear genomes. This approach has also been used to find novel nuclear-encoded genes that likely play an important role in mitochondrial dynamics, as such genes can show similar ERCs as nuOXPHOS genes (Yan et al. 2019; Williams et al. 2019). The lack of high-quality genomic data in many invertebrates is currently a hindrance to such studies, but will likely not be so for long.

Because CII, the only OXPHOS complex exclusively formed by nuclear-encoded subunits, did not show a strong ERC with either mtOXPHOS or nuOXPHOS subunits of chimeric complexes (Figure 4), it is most likely that mitonuclear coevolution, not relaxation of constraints for OXPHOS function in general, is driving the strong ERC between mt and nuOXPHOS genes. Supporting this, ERCs were generally stronger within a complex compared to across complexes (Figure 4). All of these observations are consistent with mitonuclear coevolution in bivalves.

Our finding that mtOXPHOS rates are correlated with nuOXPHOS rates, but not those of other nuclear genes, has interesting ramifications on bivalve phylogenetic inference. For example, the ERC may be driving the pattern of mitonuclear concordance described above. If mt and nuOXPHOS genes show similar rates of protein evolution compared to other nuclear genes, then long-branch attraction issues may affect nuclear phylogenies based on different genes differently. Such a scenario is consistent with the disagreement at deep nodes between mitochondrial and previous nuclear phylogenies. More focused analyses involving other nuclear markers and finer phylogenetic methods might be worthwhile.

Limited Signals of Nuclear Compensation in Bivalves

Many studies of bilaterian animals show that dN/dS ratios are extremely low in mitochondrial genes, despite low effective population size and higher mutation rate, suggesting strong selective constraints acting on mtOXPHOS subunits (distribution of values from 1st to 3rd quantile < 0.05; see for instance: Nabholz et al. 2013; Popadin et al. 2013; Havird and Sloan 2016). In the present study, dN/dS for mtOXPHOS subunits was an order of magnitude higher than those previously reported for most Metazoa, with a median value of 0.2241. This value is consistent with recent work that compared the congeneric bivalve species *R. philippinarum* and *R. decussatus* (Iannello et al. 2019) and with values calculated among mitochondrial genomes across Bivalvia (Plazzi et al. 2016). A possible explanation could be that lower metabolic needs of bivalves (due to a sedentary lifestyle) result in relaxed selection on mtOXPHOS proteins (as observed for loss of flight: Mitterboeck and Adamowicz 2013; and swimming performances: Strohm et al. 2013). Another, mutually non-exclusive, hypothesis could be that adaptations to stress tolerance (Sokolova 2018; Sokolova et al. 2019) increased the robustness of the OXPHOS system to nonsynonymous substitutions without relevant consequences in terms of fitness.

Although the biological reasons for high dN/dS in bivalve mtOXPHOS proteins are unclear, they may provide insights into mito-nuclear coevolutionary dynamics. According to the "nuclear compensation hypothesis", nuOXPHOS subunits are the prime sites for compensatory changes that maintain proper functioning of OXPHOS complexes in the face of deleterious mitochondrial mutations (Dowling et al. 2008; Gershoni et al. 2010; Osada and Akashi 2012; Havird and Sloan 2016). Some support for this hypothesis was provided in our study by the entirely nuclear-encoded CII, which had significantly lower dN/dS compared to all other, mt-interacting nuclear components (Figure 5B, Supplementary Figure 5). Similarly, the dN/dS of "non-contact" nuOXPHOS subunits was significantly lower than those of "contact" subunits (Figure 5A), which are the most obvious sites for potential compensatory changes (although this may be an oversimplification).

However, in bivalves overall dN/dS of “contact” nuOXPHOS subunits was not elevated compared to mtOXPHOS subunits (Figure 5A), unlike in most animals (Nabholz et al. 2013; Havird and Sloan 2016). Under nuclear compensation, it is generally assumed that dN/dS should be elevated in nuOXPHOS subunits, reflecting positive selection for compensatory changes. When considering each complex separately, this signal is not uniform (Figure 5B). CIV does show the expected trend under nuclear compensation of nuclear “contact” proteins that appear to evolve significantly faster than mitochondrial ones. However, CIV is constituted by some of the slowest-evolving mtOXPHOS subunits. Therefore, nuclear compensation might be expected to show the weakest signal in CIV. One possibility is that each complex and each set of subunits are undergoing different evolutionary dynamics that are driven by specific selective pressures, rather than all complexes being primarily shaped by coevolutionary constraints (see for example: Zhang and Broughton 2013; Iannello et al. 2019). Future studies might benefit by examining each complex and each subunit separately to reveal different selective pressures associated with different functional constraints. Another possibility is that the elevated dN/dS ratios in mitochondria- and nuclear-encoded “contact” subunits could be due to different reasons. Relaxed selection on mitochondrial genes coupled with positive selection on nuclear-encoded “contact” subunits could result in similarly high dN/dS ratios and would be consistent with nuclear compensation. Phylogenetic and population genetic tools to explicitly test for positive selection may be useful in exploring this possibility (Wertheim et al. 2015; Havird et al. 2017).

We also examined signatures of nuclear compensation in site-specific signals of positive selection, predicting that “contact” nuOXPHOS subunits should be enhanced for positive selection. However, out of 8 nuOXPHOS subunits in which positively selected sites were inferred, only NDUFB2 and NDUFS2 were predicted to physically contact mitochondria-encoded subunits. All other proteins represent key-subunits involved in catalysis and are located in regions of the complexes that are distant to mitochondria-encoded proteins. We acknowledge that in order to be tied by coevolutionary constraints, residues do not necessarily need to be in physical contact, since perturbations in the tertiary structure due to an amino acid mutation can compromise stability also in distant residues. However, two subunits of CII were among the putative positively selected sequences and comparable numbers of positively selected sites were found in mtOXPHOS subunits, further reducing the possibility that these results were a reflection of compensatory nuclear evolution. It could be possible that these signals of positive selection were the result of false positives due to the higher rates of sequence conservation of these proteins (see Supplementary Figure 4; Anisimova et al. 2002). These sites may simply represent residues under loose purifying selection due to their exposition in the mitochondrial matrix (therefore not involved in catalysis nor structural conformation). Regardless, our site-specific analyses do not support nuclear compensation, as in the dN/dS analyses.

Heterogeneity of Mitonuclear Evolutionary Dynamics Across Metazoa

The extent of nuclear compensatory evolution may vary among taxa. For example, in corals, dipterans, and some fungi mitochondrial and nuclear dS values are fairly similar (Havird and Sloan 2016), while Vertebrata show the highest values of mutation rate ratios between mitochondrial and nuclear genomes calculated so far (up to an average ratio per gene of 32.5 in primates: Allio et al. 2017). In our samples, mitochondrial dS values are not high on average (median is ~0.3; Supplementary Figure 7), however, these values are still noticeably higher than the nuclear ones, and the same is observed for nonsynonymous substitution rates (dN; Supplementary Figure 7). Precisely, the ratio of dS between mtOXPHOS and nuOXPHOS genes in our samples was ~2.5 (ratio between the two medians), similar to the ratio recently calculated in Bivalvia based on comparisons between mutation rates of mitochondrial genes and 398 nuclear non-mitochondria-interacting genes (median = 1.8; Allio et al. 2017). Under these conditions, we should nevertheless expect relatively higher nuclear dN/dS for our dataset under nuclear compensation (like observed in fast-mutating mtDNA taxa; Havird and Sloan 2016), but that is not the case. In other words, the high mitochondrial dN/dS observed in bivalves is not likely due entirely to a low mitochondrial mutation rate, but also due to increased rates of nonsynonymous fixations.

In our opinion, there is an important caveat to comparing dN/dS values in different genomes that may have widely differing mutation rates. Correlation between the ratio of mito-nuclear dN/dS and mito-nuclear dS observed in Havird and Sloan (2016) may have been misleading, since one of the variables is nested within the other and would automatically be expected to result in a negative correlation. Havird and Sloan attempted to control for this by examining genes without mitochondrial interactions as a control, which showed different patterns than nuOXPHOS genes. If nuclear compensation is predominantly responsible for the types of correlations observed in Havird and Sloan (2016), then amino acid substitution rate (dN) in the nuclear genes should be driving the trend. However, by reanalysing the Havird and Sloan (2016) dataset (one of the few works with a wide phylogenetic sampling across eukaryotes), we found that the mito-nuclear dN/dS ratio is only mildly correlated with the mito-nuclear dN ratio, and the correlation is driven mainly by the plant/animal dichotomy (Supplementary Figure 8). Moreover, when considering also Bivalvia values as calculated in the present study, the correlation is even weaker (Supplementary Figure 8). Other meaningful correlations, like mitochondrial dN/dS against nuclear dN/dS, or mitochondrial dN against nuclear dN, are not significant (neither excluding nor including bivalves; Supplementary Figure 9), even though they may represent more direct predictions of nuclear compensation (all correlation tests were performed with R on the dataset of Havird and Sloan 2016; data in Supplementary Material). Therefore, while mitonuclear coevolution may drive some of the observed differences between dN/dS in mt and nuOXPHOS genes in many metazoans, the large difference in underlying mutation rates between the two genomes certainly also contributes. In a case where mutation rate is high (e.g., high mitochondrial dS) but purifying selection is very strong, the need for compensation may not be high, since few protein residues actually change (e.g., low mitochondrial dN). Disentangling such nuances of dN/dS analyses should be a goal of future work.

Is DUI compatible with Nuclear Compensation?

Mitonuclear coevolution is particularly interesting in bivalves because of the frequent occurrence of DUI. In the present work, given the low representativeness of DUI species in online database, and given the additional difficulties in extracting both F- and M-type mtDNAs within a species, we could not include more DUI-specific analyses. The only signals we could get from the present sampling were represented by a handful of genes for which a specific tagging of DUI branches resulted in a better fitting Codeml model, and a slight lowering in the mtOXPHOS-nuOXPHOS branch length correlation when considering M-type mitochondria-encoded subunits for the 4 available species (almost exclusively driven by the two Unionid species; data not shown).

However, the DUI system presents some interesting considerations for mitonuclear coevolution and nuclear compensation. In DUI species, two highly divergent mitochondrial genomes have to cofunction with the same nuclear background. This introduces a potential challenge for the nuclear compensation theory because nuclear changes must offset changes happening in the two lineages of mitochondrial genes. For instance, if a mutation arises in an F-type subunit, the nuclear compensatory mutation might disrupt the co-assembly with the corresponding M-type subunit, lowering the efficiency of M-type mitochondria. However, M-type genomes, despite being usually rare in somatic tissues (but with exceptions, see Ghiselli et al. 2011), are still functionally important, since the whole male germline relies exclusively on them (Ghiselli et al. 2013; Milani and Ghiselli 2015).

One explanation for maintenance of DUI along with nuclear compensation could be the presence of two separate sets of nuOXPHOS genes that underwent duplication and evolved sex-specific expression. Such male-biased nuOXPHOS orthologues are common in mammals and *Drosophila* (Gallach et al. 2010; Eslamieh et al. 2017; Havird and McConie 2019). While this explanation cannot be completely excluded and future studies should examine it more thoroughly, no clues of duplicated sets of nuOXPHOS genes have been found so far in DUI species (Maeda et al. 2021), and we found only a single transcript per gene in all DUI species in the present study, with the exception of *M. edulis* COX4. Another possibility is sex-specific splice variants or sex-specific nuclear-encoded OXPHOS expression, which has been found in humans (Barshad et al. 2018).

A second explanation for the stable presence of DUI could lie in mitochondrial compensatory evolution, an underexplored version of mitonuclear coevolution. In such scenario, an amino acid change in a nuclear gene could be independently compensated in both M- and F-type mitochondrial genomes. The fact that these two highly divergent lineages have been kept evolutionarily stable for millions of years without disrupting respiratory capacity may be explained by considering the “mitochondrial compensation hypothesis” as the primary coevolutionary force. The production of more DUI-specific data in the future will allow us to properly address such questions.

Considerations on the Directions of Compensatory Mitonuclear Coevolution

Others have highlighted that mitonuclear coevolution could take many forms and deleterious-compensatory changes are only one class (Sloan et al. 2017). The nuclear compensation hypothesis has been favoured because classic evolutionary theory suggests non-recombining genomes such as mitochondrial genomes are likely to suffer from mutational meltdown (Lynch 1996; Lynch and Blanchard 1998; Neiman and Taylor 2009). Both empirical and modelling work has challenged this assumption (Cooper et al. 2015; Christie and Beekman 2017) and the assumption that mitochondrial genomes never recombine is also being undermined (Havird et al. 2019).

Mitochondrial genomes usually mutate faster and many variants of mtDNA are constitutively present in a heteroplasmic state (Burr et al. 2018). In the heteroplasmic pool, there might be some mtDNA copies that present a compensatory mutation for a novel amino acid change that occurred in a nuclear subunit. In this case, mitochondria that contain higher amounts of this “compensatory” mtDNA would present better functioning OXPHOS complexes with respect to the wild-type ones. Such mitochondria would have higher fitness than the others and might eventually be fixed (Milani and Ghiselli 2015; Burr et al. 2018; Zhang et al. 2018). The mechanisms that allow this selection are yet to be clarified, however the fact that better-performing mtDNA variants are favourably transmitted (Wilding et al. 2001; Zhou et al. 2010; Ghiselli et al. 2013; Hill et al. 2014; Milani 2015; Milani and Ghiselli 2015; Tworzydlo et al. 2016; Bilinski et al. 2017; Marlow 2017) could represent a coherent mechanism for mitochondrial compensation of nuOXPHOS mutations in very short evolutionary times.

Referring to this interpretation, it should be noted that almost all observations previously associated and explained in terms of nuclear compensation could be equally explained as mitochondrial compensations. For example, the fact that nuOXPHOS genes have higher dN/dS than nuclear non-OXPHOS genes and nuOXPHOS genes without mitochondrial counterparts (Havird and Sloan 2016; Havird et al. 2017; Li et al. 2017; Yan et al. 2019) could be due to mitochondrial compensation. NuOXPHOS genes can indeed be more variable because they can be efficiently compensated by a fast-mutating mitochondrial genome. When no compensation is possible, a structural deleterious mutation should simply be selected against. The same holds true for nuclear-encoded ribosomal proteins that form mitochondrial ribosomes (Barreto and Burton 2013; Sloan et al. 2014; Weng et al. 2016) and for aminoacyl-tRNA-synthetases that act on mt-tRNAs (Adrion et al. 2016). Also, many site-specific coevolutionary signals do not specifically favour nuclear compensation because they lack temporal data that could discern the order of appearance of the mutations (*inter alia* Gershoni et al. 2010, 2014; Levin and Mishmar 2017).

Little direct evidence supports nuclear compensation in contrast to other forms of mitonuclear coevolution, with one notable exception being the observation that nuclear changes tended to occur later in time than mitochondrial ones at contact residues in primates (Osada and Akashi 2012). However, a recent study by Wernick et al. (2019), showed *in vivo* evidence of direct mitochondrial compensation in *Caenorhabditis elegans*. In *gas-1* mutated lines, they directly observed functional recovery of OXPHOS efficiency through 60 generations in populations under food competition driven by novel mutations in *nadh1* and *nadh6* genes, which are mitochondria-encoded subunits in contact with the nuclear-encoded *gas-1*. It is therefore possible that in some cases the mitochondrial genome is responsible for compensatory mutations. Future studies focusing on specific residues and the temporal order of changes are needed.

Conclusions

Overall, a clear signal of mitonuclear coevolution in bivalves emerges from our data. Both the phylogenetic analysis and the ERC analyses showed strong evidence of shared evolutionary trajectories for mtOXPHOS and nuOXPHOS subunits in contrast to nuclear genes that do not interact with mitochondria. However, mitochondrial dN/dS in our samples were almost an order of magnitude higher than previously recorded bilaterian data and similar to nuclear dN/dS ratios, calling into question the idea of nuclear compensation as the driving force of mitonuclear coevolution in bivalves. Similar results were obtained in previous analyses of bivalves (Iannello et al. 2019). However, "contact" nuOXPHOS subunits displayed higher rates of evolution than "non-contact" and non-chimeric nuclear proteins, again supporting a general observation of mitonuclear coevolution. No site-specific signal of accelerated compensatory evolution was found in any of the nuclear OXPHOS subunits. Overall, support for nuclear compensation as the specific form of mitonuclear coevolution was scarce. This pattern is in contrast to other metazoans, possibly due to different reasons, including relaxed selection on OXPHOS proteins in sedentary living bivalves, increased selection on stress-tolerance pathways, or a combination of these factors. Examining a diverse sample of bivalve taxa we extend the evidence for mitonuclear coevolution to a novel taxonomic group, but question the ubiquity of the nuclear compensation hypothesis.

Materials and Methods

Dataset

We downloaded the RNA-Seq raw reads for a total of 40 bivalve species from the Short Read Archive (SRA) of NCBI (www.ncbi.nlm.gov/sra; Supplementary Table 1), trying to evenly represent the biodiversity of the class. When DUI species were considered, we downloaded reads from both sexes, in order to retrieve both mitochondrial genomes.

We removed sequencing primers and filtered out low-quality and unpaired reads using Trimmomatic v0.36 (Bolger et al. 2014) with the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:25:33 MINLEN:75. Transcriptomes were then assembled *de novo* using Trinity-v2.4.0 (Haas et al. 2013) with default parameters. To assess quality and completeness of the transcriptomes, we used BUSCO v2 on the Metazoa core-orthologue set (Simão et al. 2015), as implemented in the gVolante website (www.gvolante.riken.jp/analysis.html; Nishimura et al. 2017). We filtered the transcriptomes through a DIAMOND v0.9.19.120 (Buchfink et al. 2015) search against NCBI non-redundant protein database (nr), retaining only those transcripts for which the best hit was against a lophotrochozoan.

OXPHOS Subunits Annotation

MtOXPHOS transcripts were identified with a BLASTX (BLAST v2.6.0+; Camacho et al. 2009) search of each transcriptome against a custom database containing all molluscan mtOXPHOS protein coding genes (PCGs) (downloaded from NCBI). We then manually extracted Open Reading Frames (ORFs) using the NCBI ORFfinder online tool (www.ncbi.nlm.nih.gov/orffinder), validating the results with a BLASTP against nr. When both F- and M-type mitochondrial gene products were annotated in DUI species, we considered them as separate OTUs throughout the whole analysis.

ORFs of nuOXPHOS subunits were retrieved using the Findorf tool (Krasileva et al. 2013), that uses a BLASTX search against a user-defined database, and a HMMER (Mistry et al. 2013) search against the Pfam database 30.0 (Finn et al. 2016). To build the user-defined BLAST database, we downloaded nuOXPHOS protein sequences of 7 reference species from the KEGG database (https://www.genome.jp/kegg/; Kanehisa and Goto 2000): *C. gigas, Octopus bimaculoides, Lottia gigantea, Helobdella robusta, C. elegans, D. melanogaster*, and *H. sapiens*. The complete set consisted of 78 subunits (38 of CI, 4 of CII, 9 of CIII, 14 of CIV, and 13 of CV; see Figure 1). We implemented the annotation with the PSI-BLAST tool, that consists of a series of consecutive BLASTP iterations using the protein sequences positively annotated with the Findorf tool as databases to complete the annotation in the species with missing genes. We then removed from nuOXPHOS subunits the Mitochondrial Targeting Signal (MTS) since they do not participate in the coevolutionary dynamics of the mito-nuclear OXPHOS complexes and are subjected to rather different evolutionary forces (e.g. ligand-receptor interactions). To predict mitochondrial processing peptidase (MPP) cleavage sites we used MitoFates (Fukasawa et al. 2015)

Phylogenetic Inference

Two Maximum Likelihood (ML) phylogenies were inferred for the concatenated sets of mtOXPHOS and nuOXPHOS subunits. We aligned the amino acid sequences with PSICOFFEE (Floden et al. 2016) and trimmed the alignments with BMGE v1.12 (BLOSUM30 -h 0.75 -b 3; Criscuolo and

Gribaldo 2010). We inferred partitions and best-fitting models with PartitionFinderProtein (Lanfear et al. 2017). ML trees were built with RAxML v8.2.11 (Stamatakis 2014) with 1,000 bootstrap replicates, forcing bivalve monophyly, with four non-bivalve molluscs as outgroups. Nodes with a bootstrap support value lower than 0.7 were collapsed.

Evolutionary Rate Correlations

We also examined evolutionary rate correlations (ERC; a useful test to investigate protein coevolutionary dynamics, see: de Juan et al. 2013; Williams et al. 2019; Yan et al. 2019) between the mt and nuOXPHOS proteins. We built a species tree and optimized the branch lengths of the concatenated alignments with RAxML v8.2.11 (Stamatakis 2014). The species tree of our sample species was built manually using data from the literature (see the tree in Figure 1). We mostly referred to the phylotranscriptomic analysis of Gonzalez and colleagues (2015), considering the genus or the family for species not present in that work. The inner relationships among the three Unionidae species considered here (*Cristaria plicata, Lampsilis cardium, Hyriopsis cumingii*) could not be solved for lack of confident literature, and we keep them as a polytomy. Relationships within Pteriomorphia were based on the phylogenomic work of Lemer and colleagues (2016).

A set of random nuclear proteins was used as control for the ERC; for this purpose we used Proteinortho v6.0.7 (Lechner et al. 2011) to obtain ortholog transcripts from the 31 bivalve transcriptomes of our study. We selected 24 orthologue clusters from the output (maximizing the species representation) for a total of 605 transcripts (139 missing sequences). We extracted ORFs with TransDecoder (https://github.com/TransDecoder/TransDecoder), and through a BLASTP search against the nr database we ensured that no mitochondria-interacting proteins were included in these clusters. Alignments, trimming, partitioning, branch length optimization on the species tree, and distance to the root calculations were performed as for the OXPHOS proteins. We then performed correlation tests (cor.test function in R) on the distances to the root (patristic method of distRoot function in R, adephylo package) of every species in the three different sets of proteins as a proxy for coevolutionary dynamics. Correlating distributions of distances to the root introduces non-independence among within-distribution values because of shared branches, and this could bias the calculations. For this reason, we also checked for correlation between the lengths of the terminal branches (i.e. species-specific), founding no differences.

Another possible bias that may affect ERC is the potential non-random representativeness of the 24 nuclear control orthologue clusters. To test this, we randomly divided the cluster in two subsets of 12 proteins and tested for ERC between the two. This was performed 1,000 times to obtain a median correlation coefficient and its confidence intervals for each ERC. Moreover, we divided both the mtOXPHOS and the nuOXPHOS for the branch lengths of both orthologues subsets to check if the correlation held after a normalization to control for variation among species in overall rates of nuclear evolution (normalization performed for all 1,000 iterations). To check for more specific coevolutionary signals, we calculated branch lengths for separate datasets of both mitochondria- and nuclear-encoded subunits of each complex and tested for correlations between the different components.

Rates of Protein Evolution

For the analyses on the rates of protein evolution, we followed the same alignment procedure for the phylogenetic analyses. We optimized branch lengths of the species tree (see above) for each alignment of our analyses with RAxML v8.2.11 (Stamatakis 2014). Best fitting models for RAxML were inferred with ProtTest v3.4 (Darriba et al. 2011). We calculated for each backtranslated alignment dN/dS using Codeml (PAML v4.9 package; Yang 2007; Supplementary Figure 3).

Each alignment was tested for a free-ratio model of dN/dS calculation over the tree (each branch associated to a different value, i.e. branch-model 1; model = 1) against a uniform-rate model (a single value averaged for all branches, i.e. branch-model 0; model = 0). The best-fitting model was estimated through Likelihood Ratio Tests (LRTs), comparing Log-likelihood values for each model (maximized over 6 replicates). In the cases where branch-model 1 was the best-fitting model, we pooled together dN/dS values for all branches of each subunit tree, therefore associating a distribution to each gene product, rather than a single value. To be able to compare the single-ω subunits with the others, we replicated the single dN/dS value for all the branches of their trees, therefore equally weighing the two sets of subunits in the overall distribution.

We divided the nuOXPHOS subunits in two clusters: those predicted to be in physical contact to mitochondria-encoded subunits and those without any supposed direct interaction with mtOXPHOS proteins (Complex IV: Richter and Ludwig 2003; Complex V: Jonckheere et al. 2012; Complex I: Zhu et al. 2016; Complex II: Amporndanai et al. 2018). Statistical group analyses were conducted with Wilcoxon-Mann-Whitney and Dunn tests (with Bonferroni correction) as implemented in R v3.4.4. Zero values of dN or dS, that resulted in calculations of dN/dS of either 0 or 999 in Codeml, were excluded.

In our dataset we included 7 DUI species, that are known to possess two different mitochondrial genomes that are maintained separately by sex-specific segregation (see Introduction). Since we pooled all dN/dS values of the tree together, we tested whether the DUI species biased the overall signal, especially for the subunits where the free-ratio model was better than the single-ω model. In order to do so, we performed LRTs between single-ω model and the branch-specific model, that allows to tag different branches or clades for which a specific dN/dS is calculated (in this case we tagged the private branch of each DUI species – or the whole clade in the case of Unionida). We indeed found a handful of cases for which the branch-specific model was better. Such results were however confined to few genes of the dataset and did not allow us to consider the DUI phenomenon as a source of bias for the analysis. To double check, we removed these genes and reran all the analyses, observing no change from any result.

Signatures of Positive Selection

We also used Codeml to investigate the site-specific evolutionary rate of gene products (graphic summary: Supplementary Figure 6). To test whether a model considering different dN/dS for different sites fit the data better than one implementing a uniform rate, we tested Log-likelihood values (maximum values over 6 calculation replicates) of M0 (single dN/dS; NSsites = 0) and M3 (n categories of dN/dS: 5 in our case; NSsites = 3) with LRTs. When M3 was the best model, we tested for the presence of positive selection comparing two pairs of models. Each pair consisted in a model that included parameters admitting positively selected sites, and another that did not (the null model; Yang 2007): they were M1a (variable selective pressure but no positive selection; NSsites = 1) vs M2a (M1a plus positive selection; NSsites = 2) and M7 (beta distributed variable selective pressure; NSsites = 7) vs M8 (M7 plus positive selection; NSsites = 8). When both the models that included sites with dN/dS>1 were the best, we performed additional tests to evaluate whether we could actually consider positive selection as a possibility. We tested M8 against the MEC model, which takes into account the weight of each amino acid replacement (Doron-Faigenboim and Pupko 2007) in terms of radical and conservative modifications based on empirical replacement probability matrices (calculation performed on the Selecton server; Stern et al. 2007). Those models are not nested within each other, therefore an LRT was not possible. Hence, we compared Akaike Information Criteria scores in order to evaluate the best model. When M8 was the best, we considered the sites under positive selection as predicted by the Bayes Empirical Bayes (BEB) method as implemented in Codeml.

These results were then compared with annotated ligand and catalytic sites from the literature. In detail, we compared the protein sequences of *C. angulata* (as annotated in our dataset) with the functional sites as predicted in *C. gigas* (NCBI protein database), or with annotated sites in *H. sapiens* when such information was not available for any *Crassostrea* species. Sites under putative positive selection were plotted on the tertiary structures of the *C. angulata* proteins. Structural conformations were predicted on the I-TASSER server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/; Zhang 2008) with a C-score cut-off of 0 (C-score is typically in the range [-5,2], with higher values representing more confident model predictions). Structural alignment against known structures of the OXPHOS complexes (downloaded from the Protein Data Bank archives, https://www.rcsb.org) were performed in order to visualize the sites of interest in the context of the quaternary structure (CI: *H. sapiens*, 10.2210/pdb5XTD/pdb; CII: *Escherichia coli*, 10.2210/pdb1NEK/pdb; CIV and CV: *Bos taurus*, 10.2210/pdb5XDX/pdb and 10.2210/pdb5ARA/pdb, respectively).

References

Adrion JR, White PS, Montooth KL. 2016. The Roles of Compensatory Evolution and Constraint in Aminoacyl tRNA Synthetase Evolution. Mol. Biol. Evol. 33:152–161.

Allio R, Donega S, Galtier N, Nabholz B. 2017. Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. Mol. Biol. Evol. 34:2762–2772.

Amporndanai K, Johnson RM, O'Neill PM, Fishwick CWG, Jamson AH, Rawson S, Muench SP, Hasnain SS, Antonyuk SV. 2018. X-ray and cryo-EM structures of inhibitor-bound cytochrome bc1 complexes for structure-based drug discovery. IUCrJ 5:200–210.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19:950–958.

Ballard JW, Whitlock MC. 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13:729–744.

Barreto FS, Burton RS. 2013. Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. Mol. Biol. Evol. 30:310–314.

Barreto FS, Watson ET, Lima TG, Willett CS, Edmands S, Li W, Burton RS. 2018. Genomic signatures of mitonuclear coevolution across populations of Tigriopus californicus. Nat Ecol Evol 2:1250–1257.

Barshad G, Blumberg A, Cohen T, Mishmar D. 2018. Human primitive brain displays negative mitochondrial-nuclear expression correlation of respiratory genes. Genome Res 28:952-967

Bar-Yaacov D, Blumberg A, Mishmar D. 2012. Mitochondrial-nuclear co-evolution and its effects on OXPHOS activity and regulation. Biochim. Biophys. Acta 1819:1107–1111.

Bieler R, Mikkelsen PM, Collins TM, Glover EA, González VL, Graf DL, Harper EM, Healy J, Kawauchi GY, Sharma PP, et al. 2014. Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. Invertebr. Syst. 28:32.

Bilinski SM, Kloc M, Tworzydlo W. 2017. Selection of mitochondria in female germline cells: is Balbiani body implicated in this process? J. Assist. Reprod. Genet. 34:1405–1412.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Breton S, Stewart DT, Hoeh WR. 2010. Characterization of a mitochondrial ORF from the gender-associated mtDNAs of Mytilus spp. (Bivalvia: Mytilidae): identification of the "missing" ATPase 8 gene. Mar. Genomics 3:11–18.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12:59–60.

Burr SP, Pezet M, Chinnery PF. 2018. Mitochondrial DNA Heteroplasmy and Purifying Selection in the Mammalian Female Germ Line. Dev. Growth Differ. 60:21–32.

Burton RS, Barreto FS. 2012. A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? Mol. Ecol. 21:4942–4957.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Christie JR, Beekman M. 2017. Uniparental Inheritance Promotes Adaptive Evolution in Cytoplasmic Genomes. Mol. Biol. Evol. 34:677–691.

Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL. 2015. Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both Drosophila melanogaster and Homo sapiens. G3 5:2165–2176.

Levin L, Mishmar D. 2017.The genomic landscape of evolutionary convergence in mammals, birds and reptiles. Nat. Ecol. Evol. 1:0041.

Li Y, Zhang R, Liu S, Donath A, Peters RS, Ware J, Misof B, Niehuis O, Pfrender ME, Zhou X. 2017. The molecular evolutionary dynamics of oxidative phosphorylation (OXPHOS) genes in Hymenoptera. BMC Evol. Biol. 17:269.

Lopes-Lima M, Froufe E, Do VT, Ghamizi M, Mock KE, Kebapçı Ü, Klishko O, Kovitvadhi S, Kovitvadhi U, Paulo OS, et al. 2017. Phylogeny of the most species-rich freshwater bivalve family (Bivalvia: Unionida: Unionidae): Defining modern subfamilies and tribes. Mol. Phylogenet. Evol. 106:174–191.

Lynch M. 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. Mol. Biol. Evol. 13:209–220.

Lynch M, Blanchard JL. 1998. Deleterious mutation accumulation in organelle genomes. Genetica 102-103:29–39.

Maldonado E, Sunagar K, Almeida D, Vasconcelos V, Antunes A. 2014. IMPACT_S: integrated multiprogram platform to analyze and combine tests of selection. PLoS One 9:e96243.

Marlow FL. 2017. Mitochondrial matters: Mitochondrial bottlenecks, self-assembling structures, and entrapment in the female germline. Stem Cell Res. 21:178–186.

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370:20140330.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus–cytosol compartmentalization. Nature 440:41–45.

Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.

McKenzie M, Chiotis M, Pinkert CA, Trounce IA. 2003. Functional respiratory chain analyses in murid xenomitochondrial cybrids expose coevolutionary constraints of cytochrome b and nuclear subunits of complex III. Mol. Biol. Evol. 20:1117–1124.

Milani L. 2015. Mitochondrial membrane potential: a trait involved in organelle inheritance? Biol. Lett. 11(10):20150732.

Milani L, Ghiselli F. 2015. Mitochondrial activity in gametes and transmission of viable mtDNA. Biol. Direct 10:22.

Milani L, Ghiselli F. 2020. Faraway, so close. The comparative method and the potential of non-model animals in mitochondrial research. Philos. Trans. R. Soc. Lond. B Biol. Sci. 375:20190186.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41:e121.

Mitterboeck TF, Adamowicz SJ. 2013. Flight loss linked to faster molecular evolution in insects. Proc. Biol. Sci. 280:20131128.

Nabholz B, Ellegren H, Wolf JBW. 2013. High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. Mol. Biol. Evol. 30:272–284.

Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes. Proc. Biol. Sci. 276:1201–1209.

Niehuis O, Judson AK, Gadau J. 2008. Cytonuclear genic incompatibilities cause increased mortality in male F2 hybrids of Nasonia giraulti and N. vitripennis. Genetics 178:413–426.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10:210.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. Mol. Biol. Evol. 24:388–397.

Doucet-Beaupré H, Breton S, Chapman EG, Blier PU, Bogan AE, Stewart DT, Hoeh WR. 2010. Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. BMC Evol. Biol. 10:50.

Dowling DK, Friberg U, Lindell J. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. Trends Ecol. Evol. 23:546–554.

Dreyer H, Steiner G. 2006. The complete sequences and gene organisation of the mitochondrial genomes of the heterodont bivalves Acanthocardia tuberculata and Hiatella arctica--and the first record for a putative Atpase subunit 8 gene in marine bivalves. Front. Zool. 3:13.

Eslamieh M, Williford A, Betrán E. 2017. Few Nuclear-Encoded Mitochondrial Gene Duplicates Contribute to Male Germline-Specific Functions in Humans. Genome Biol. Evol. 9:2782–2790.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.

Floden EW, Tommaso PD, Chatzou M, Magis C, Notredame C, Chang J-M. 2016. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Nucleic Acids Res. 44:W339–W343.

Forni G, Puccio G, Bourguignon T, Evans T, Mantovani B, Rota-Stabelli O, Luchetti A. 2019. Complete mitochondrial genomes from transcriptomes: assessing pros and cons of data mining for assembling new mitogenomes. Sci. Rep. 9:14806.

Fukasawa Y, Tsuji J, Fu S-C, Tomii K, Horton P, Imai K. 2015. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol. Cell. Proteomics 14:1113–1126.

Gallach M, Chandrasekaran C, Betrán E. 2010. Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic conflict in Drosophila. Genome Biol. Evol. 2:835–850.

Gershoni M, Fuchs A, Shani N, Fridman Y, Corral-Debrinski M, Aharoni A, Frishman D, Mishmar D. 2010. Coevolution predicts direct interactions between mtDNA-encoded and nDNA-encoded subunits of oxidative phosphorylation complex I. J. Mol. Biol. 404:158-171.

Gershoni M, Levin L, Ovadia O, Toiw Y, Shani N, Dadon S, Barzilai N, Bergman A, Atzmon G, Wainstein J, Tsur A, Nijtsmans L, Glaser B, Mishmar D. 2014. Disrupting mitochondrial-nuclear coevolution affects OXPHOS complex I integrity and impacts human health. Genome Biol. Evol. 6(10): 2665-2680.

Ghiselli F, Maurizii MG, Reunov A, Ariño-Bassols H, Cifaldi C, Pecci A, Alexandrova Y, Bettini S, Passamonti M, Franceschini V, et al. 2019. Natural Heteroplasmy and Mitochondrial Inheritance in Bivalve Molluscs. Integr. Comp. Biol. 59:1016–1032.

Ghiselli F, Milani L, Guerra D, Chang PL, Breton S, Nuzhdin SV, Passamonti M. 2013. Structure, transcription, and variability of metazoan mitochondrial genome: perspectives from an unusual mitochondrial inheritance system. Genome Biol. Evol. 5:1535–1554.

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. Bioinformatics 33:3635–3637.

Osada N, Akashi H. 2012. Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. Mol. Biol. Evol. 29:337–346.

Plazzi F, Puccio G, Passamonti M. 2016. Comparative Large-Scale Mitogenomics Evidences Clade-Specific Evolutionary Trends in Mitochondrial DNAs of Bivalvia. Genome Biol. Evol. 8:2544–2564.

Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. 2013. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. Mol. Biol. Evol. 30:347–355.

Rand DM, Haney RA, Fry AJ. 2004. Cytonuclear coevolution: the genomics of cooperation. Trends Ecol. Evol. 19:645–653.

Richter O-MH, Ludwig B. 2003. Cytochrome c oxidase — structure, function, and physiology of a redox-driven molecular machine. In: Reviews of Physiology, Biochemistry and Pharmacology. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 47–74.

Sharma PP, González VL, Kawauchi GY, Andrade SCS, Guzmán A, Collins TM, Glover EA, Harper EM, Healy JM, Mikkelsen PM, et al. 2012. Phylogenetic analysis of four nuclear protein-encoding genes largely corroborates the traditional classification of Bivalvia (Mollusca). Mol. Phylogenet. Evol. 65:64–74.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.

Sloan DB, Havird JC, Sharbrough J. 2017. The on-again, off-again relationship between mitochondrial genomes and species boundaries. Mol. Ecol. 26:2212–2236.

Sloan DB, Triant DA, Wu M, Taylor DR. 2014. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Mol. Biol. Evol. 31:673–682.

Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, Havird JC. 2018. Cytonuclear integration and co-evolution. Nat. Rev. Genet. 19:635–648.

Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716.

Sokolova I. 2018. Mitochondrial Adaptations to Variable Environments and Their Role in Animals' Stress Tolerance. Integr. Comp. Biol. 58:519–531.

Sokolova IM, Sokolov EP, Haider F. 2019. Mitochondrial Mechanisms Underlying Tolerance to Fluctuating Oxygen Conditions: Lessons from Hypoxia-Tolerant Organisms. Integr. Comp. Biol. 59:938–952.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Ghiselli F, Milani L, Passamonti M. 2011. Strict sex-specific mtDNA segregation in the germ line of the DUI species Venerupis philippinarum (Bivalvia: Veneridae). Mol. Biol. Evol. 28:949–961.

González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, Taylor JD, Giribet G. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proc. Biol. Sci. 282:20142332.

Gray MW. 2012. Mitochondrial evolution. Cold Spring Harb. Perspect. Biol. 4:a011403.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. Science 283:1476–1481.

Gusman A, Lecomte S, Stewart DT, 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.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Havird JC, Forsythe ES, Williams AM, Werren JH, Dowling DK, Sloan DB. 2019. Selfish Mitonuclear Conflict. Curr. Biol. 29:R496–R511.

Havird JC, McConie HJ. 2019. Sexually Antagonistic Mitonuclear Coevolution in Duplicate Oxidative Phosphorylation Genes. Integr. Comp. Biol. 59:864–874.

Havird JC, Sloan DB. 2016. The Roles of Mutation, Selection, and Expression in Determining Relative Rates of Evolution in Mitochondrial versus Nuclear Genomes. Mol. Biol. Evol. 33:3042–3053.

Havird JC, Trapp P, Miller C, Bazos I, Sloan DB. 2017. Causes and consequences of rapidly evolving mtDNA in a plant lineage. Genome Biol. Evol. 9(2): 323-336

Healy TM, Burton RS. 2020. Strong selective effects of mitochondrial DNA on the nuclear genome. Proc. Natl. Acad. Sci. U. S. A. 117:6616–6621.

Hill GE. 2015. Mitonuclear Ecology. Mol. Biol. Evol. 32:1917–1927.

Hill GE. 2019. Mitonuclear Ecology. New York: Oxford University Press

Hill GE. 2020. Mitonuclear Compensatory Coevolution. Trends Genet. 36:403–414.

Hill GE, Havird JC, Sloan DB, Burton RS, Greening C, Dowling DK. 2019. Assessing the fitness consequences of mitonuclear interactions in natural populations. Biol. Rev. Camb. Philos. Soc. 94:1089–1104.

Hill JH, Chen Z, Xu H. 2014. Selective propagation of functional mitochondrial DNA during oogenesis restricts the transmission of a deleterious mitochondrial variant. Nat. Genet. 46:389–392.

Iannello M, Puccio G, Piccinini G, Passamonti M, Ghiselli F. 2019. The dynamics of mito-nuclear coevolution: A perspective from bivalve species with two different mechanisms of mitochondrial inheritance. J. Zoolog. Syst. Evol. Res. 57:534–547.

Jonckheere AI, Smeitink JAM, Rodenburg RJT. 2012. Mitochondrial ATP synthase: architecture, function and pathology. J. Inherit. Metab. Dis. 35:211–225.

de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. Nat. Rev. Genet. 14:249–261.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. 2007. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. 35:W506–W511.

Stöger I, Schrödl M. 2013. Mitogenomics does not resolve deep molluscan relationships (yet?). Mol. Phylogenet. Evol. 69:376–392.

Strohm JHT, Gwiazdowski RA, Hanner R. 2015. Fast fish face fewer mitochondrial mutations: Patterns of dN/dS across fish mitogenomes. Gene 572:27–34.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5:123–135.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Mol. Ecol. 21:3907–3930.

Tworzydlo W, Kisiel E, Jankowska W, Witwicka A, Bilinski SM. 2016. Exclusion of dysfunctional mitochondria from Balbiani body during early oogenesis of Thermobia. Cell Tissue Res. 366:191–201.

Weng M-L, Ruhlman TA, Jansen RK. 2016. Plastid-Nuclear Interaction and Accelerated Coevolution in Plastid Ribosomal Genes in Geraniaceae. Genome Biol. Evol. 8:1824–1838.

Wernick RI, Christy SF, Howe DK, Sullins JA, Ramirez JF, Sare M, Penley MJ, Morran LT, Denver DR, Estes S. 2019. Sex and Mitonuclear Adaptation in Experimental Caenorhabditis elegans Populations. Genetics 211:1045–1058.

Wertheim JO, Murrel B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol 32(3):820-832

Wilding M, Carotenuto R, Infante V, Dale B, Marino M, Di Matteo L, Campanella C. 2001. Confocal microscopy analysis of the activity of mitochondria contained within the “mitochondrial cloud” during oogenesis in Xenopus laevis. Zygote 9:347–352.

Williams AM, Friso G, van Wijk KJ, Sloan DB. 2019. Extreme variation in rates of evolution in the plastid Clp protease complex. Plant J. 98:243–259.

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA. 2003. TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics 19:671–672.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yan Z, Ye G, Werren JH. 2019. Evolutionary Rate Correlation between Mitochondrial-Encoded and Mitochondria-Associated Nuclear-Encoded Proteins in Insects. Mol. Biol. Evol. 36:1022–1036.

Zachar I, Szathmáry E. 2017. Breath-giving cooperation: critical review of origin of mitochondria hypothesis. Biol. Direct 12: 19.

Zhang F, Broughton RE. 2013. Mitochondrial-nuclear interactions: compensatory evolution or variable functional constraint among vertebrate oxidative phosphorylation genes? Genome Biol. Evol. 5:1781–1791.

Zhang H, Burr SP, Chinnery PF. 2018. The mitochondrial DNA genetic bottleneck: inheritance and beyond. Essays Biochem. 62:225–234.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

Zhou RR, Wang B, Wang J, Schatten H, Zhang YZ. 2010. Is the mitochondrial cloud the selection machinery for preferentially transmitting wild-type mtDNA between generations? Rewinding Müller’s ratchet efficiently. Curr. Genet. 56:101–107.

Zhu J, Vinothkumar KR, Hirst J. 2016. Structure of mammalian respiratory complex I. Nature 536:354–358.

Kocot KM, Cannon JT, Todt C, Citarella MT, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, Halanych KM. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452-456

Kolesnikov AA, Gerasimov ES. 2012. Diversity of mitochondrial genome organization. Biochemistry 77:1424–1435.

Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, Soria M, Wang S, IWGS Consortium, Akhunov E, et al. 2013. Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol. 14:R66.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929–934.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Mol. Biol. Evol. 34:772–773.

Zouros E. 2013. Biparental Inheritance Through Uniparental Transmission: The Doubly Uniparental Inheritance (DUI) of Mitochondrial DNA. Evol. Biol. 40:1–31.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124.

Lemer S, González VL, Bieler R, Giribet G. 2016. Cementing mussels to oysters in the pteriomorphian tree: a phylogenomic approach. Proceedings of the Royal Society B: Biological Sciences 283:20160857.