

Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze della Terra, della Vita e dell' Ambiente

Ciclo XXX

Settore Concorsuale: 05/B1 – ZOOLOGIA E ANTROPOLOGIA

Settore Scientifico Disciplinare: BIO/05 - ZOOLOGIA

**A comparative transcriptomic study on the evolution of
nuclear and mitochondrial genes in bivalves**

Presentata da: Dott. Mariangela Iannello

Coordinatore Dottorato

Prof. Giulio Viola

Supervisore

Prof. Marco Passamonti

Co-supervisore

Dott. Fabrizio Ghiselli

Esame finale anno 2018

Abstract

With more than 100.000 extant species, Mollusca is the second Phylum for number of species after arthropods. Molluscs are abundant in most marine and terrestrial environments and some species have adapted to live in extreme conditions. Also, this taxon shows a great diversity in term of morphology, size, complexity and behavioral repertoires. All these features make mollusc species excellent candidates for studying evolution. Nevertheless, few comparative genomic or transcriptomic works are present in literature and most of the biological questions investigated so far remain unexplored in this Phylum. In addition, most of the bioinformatics tools required to analyze High Throughput Sequencing (HTS) data are optimized for model species, making the investigation of non-model organisms far to be straightforward.

During my PhD, my research activity was twofold: I first developed a pipeline specifically designed for the annotation of transcriptomes in non-model animals; then I used RNA-Seq data to investigate transcriptomes from mature gonads of *R. decussatus* and *R. philippinarum* (Bivalvia, Veneridae), focusing my analyses on the evolution of sex-biased genes and on the co-evolution of mitochondrial and nuclear genomes.

Summary

Introduction	2
High Throughput Sequencing for studying evolution	2
The importance of animal non-model species	5
Why Mollusca?	8
An unusual mechanism of mitochondrial inheritance	12
Aim of this thesis	16
Short summaries	18
Chapter 1: A transcriptome annotation pipeline for non-model organisms	22
1) Contaminant Filtering	26
2) ORF prediction	28
3) High-quality amino acid sequence annotation	31
4) Identification of orthologs and paralogs	33
5) Nucleotide annotation	35
6) Low-quality amino acid sequence annotation	37
Chapter 2: The complete mitochondrial genome of the grooved carpet shell, <i>Ruditapes decussatus</i> (Bivalvia, Veneridae)	40
Chapter 3: Comparative transcriptomics in two bivalve species offers different perspectives on the evolution of sex-biased genes	79
Chapter 4: No evidence for nuclear compensation hypothesis in species with different mechanisms of mitochondrial inheritance	133
Conclusions	181
References	184

Introduction

High Throughput Sequencing for studying evolution

The development of High-Throughput Sequencing (HTS) changed completely the way of doing biological research. Since the first human genome was completed in 2001, with an estimated cost of 2.7 billion dollars and 10 years of work, large technological improvements have been made. Today, HTS have drastically reduced costs and times: different platforms have been developed to optimize the sequencing and the quality and length of the reads have been rapidly increasing (Table 1) (da Fonseca et al. 2016). As a consequence, obtaining complete genomes and transcriptomes is now easier for scientists and an unprecedented volume of data is available on public databases, such as the Sequence Read Archive (SRA) or the Transcriptome Shotgun Assembly (TSA) (<http://www.ncbi.nlm.nih.gov>). In addition, many projects aim at increasing the number of taxa sequenced: good examples are the Genome 10K Project, established for sequencing 10,000 vertebrate genomes (Koepfli et al. 2015).

The availability of this huge amount of data allows to investigate virtually any biological question: HTS are indeed largely used in ecology, genetics, conservation biology, and biomedical fields. Moreover, the possibility to compare genomes or transcriptomes had a profound impact on evolutionary biology. Choosing the proper sequencing approach, it is possible to investigate biological questions at any evolutionary time scale. As shown in figure 1, different HTS methods are applied at distinct evolutionary times, providing insights from deep level relationship in Metazoa (see for example Smith et al. 2011), to microevolutionary mechanisms, such as in closely-related species or populations (Parker et al. 2017).

Genomes and transcriptomes are therefore a crucial resource for studying evolution, and they are largely used for phylogenetic analyses, to investigate development,

genome structures, alternative splicing, regulatory networks, protein family expansion/contraction and gene gain/loss across taxa.

Sequencing technology/Platform	Detection method	Library types	Maximum read length (bp)	Reds per run (maximum)	Error rate (approximate)	Pros	Cons
454/GS FLX titanium XL+ 454/GS Junior Systems	Pyrophosphate detection	Single end/ Paired end	1000 800	1,000,000 100,000	0.2%	Medium size reads. Errors are well characterized	Will be discontinued in 2016. Inaccurate homopolymer detection
Illumina-Solexa/HiSeq 4000 Illumina-Solexa/MySeq	Fluorescence, reversible terminators	Single end/ Paired end	2 × 150 2 × 300	5,000,000,000 25,000,000	0.2–0.8%	Widely used. Flexible library preparation methods. High-throughput well suited for resequencing projects. Good characterization of biases	Emulsion PCR. Short reads. Not optimal for de novo assembly.
Life Technologies/SOLID	Fluorescence di-base probes	Single end/ Paired end	1 × 75/2 × 50	1,400,000,000	0.01%	Second most used	Color space not supported by many mappers. Short reads. Emulsion PCR.
Life Technologies/Ion Torrent	Hydrogen ion (pH) sensor	Single end/ Paired end	400	5,500,000	1.8%	Second highest throughput. Each base is read twice, thus decreasing the error rate. Short running times.	Bias against AT-rich regions
PacBio RS II/SMRT	Fluorescence phospho-linked nucleotides	Single end	20,000	55,000	13%	Longest reads. Good for improving de novo assemblies. Single molecule sequencing. Long reads. No GC bias	Inaccurate homopolymer detection Emulsion PCR. Low throughput. High cost-throughput ratio. High error rates
Oxford Nanopore/MintION	Electrical sensing	Single end	2000	60,000	30%	Portable. Scalable. Real-time data. Single molecule. It is possible to read both strands of the DNA sequence.	High error rates. Quality scores only defined by the quality of the alignment to a reference sequence. Existing mapping and assembly software do not deal with long and high error reads.

Table 1. High Throughput Sequencing technologies (da Fonseca et al. 2016).

Also, by analyzing the rate of protein evolution, typically calculated as rate of nonsynonymous to synonymous substitution (dN/dS), it is possible to detect rapidly evolving genes, and estimate the contribution of natural selection and genetic drift in the evolution of different lineages. For example, the calculation of dN/dS from

thousands of genes in mammals has shown that the mean rate of protein evolution ranges between 0.10 and 0.25 (Chimpanzee Sequencing and Analysis Consortium 2005), and that about 75-90% of nonsynonymous substitution are removed by purifying selection.

A method that represented a major breakthrough is RNA-Seq (Mortazavi et al. 2008), a powerful approach to investigate biological issues also in species where a genome reference is not available, providing information about both transcript sequence and abundance. This has great impact on the study of evolution, since it is commonly known that changes in spatio-temporal gene expression play a crucial role in adaptive evolution and are responsible for divergence in development (Harrison et al. 2012). In fact, orthologous genes show large differences in expression even between closely related species. For example, most of the phenotypic differences between human and chimpanzees are not explained by changes in proteins but rather by differences in gene regulation (Romero et al. 2012). The same way, males and females share almost the same genome, and the differences in terms of sexual dimorphism and behavior are the result of genes that are preferentially expressed in one of the other sex (Ellegren & Parsch 2007). For a detailed discussion about this topic see Chapter 3.

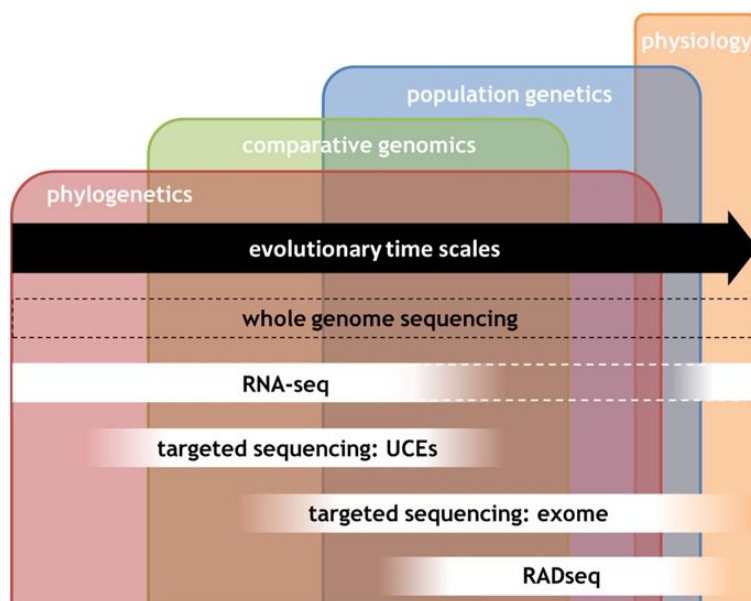


Figure 1. Application of different high-throughput sequencing methods to different evolutionary time scales (da Fonseca et al. 2016).

The importance of animal non-model species

While for some species a large amount of sequencing data is available in the public databases, there is a severe lack of data for a wide range of taxa. Within vertebrates, 50% of primate families have a genome reference, and mammals in general are well characterized; on the contrary, few data are available for reptiles and amphibian, while birds have been sequenced more extensively just very recently (Ellegren 2014). On the other hand, outside vertebrates, sequencing data are far from being representative of animal diversity: there are no reference genomes for about half of the insect orders, and much less data are available for other invertebrates (see for example Dunn & Ryan 2015). Therefore, the list of sequenced animal taxa is greatly biased toward model species, vertebrates, domesticated species as well as species used for agriculture purposes. This lack of information precludes our capabilities to infer many biological questions and, so far, the study of evolution has been based on the investigation of a limited number of species, resulting often in a generalization of the studied mechanisms and in a limitation of our comprehension about the patterns of evolution. Indeed, works on non-model organisms often reveal lineage-specific features such as, for instance, adaptations, gene gain/loss, and gene neofunctionalization. For example, the sequencing of the Pacific sea gooseberry (*Pleurobrachia bachei*) genome, with other 10 ctenophore transcriptomes, revealed the absence of HOX genes as well as classical neurotransmitter pathways. This evidence suggested that the ctenophore neural system evolved independently from that in other animals (Moroz et al. 2014). Similarly, the genome of the African coelacanth, *Latimeria chalumnae*, revealed changes in genes and regulatory element involved in immunity and development of fins, tail, ear, brain and olfaction, providing insights into vertebrate evolution and water-to-land transition. Table 2 reports some key findings obtained from sequencing of non-model species. In order to have a complete landscape of the evolution, it is necessary to fill the gap with sequencing data from the least represented taxa. The Global Invertebrate Genomics Alliance (GIGA) is moving in this direction (GIGA Community of Scientists et al. 2014;

Voolstra et al. 2017), aiming to obtain 7000 between genomes and transcriptomes of non-model invertebrate species.

On the other hand, it must be highlighted that working on non-model species is often far to be straightforward. In order to analyze the large amount of data coming from HTS, we need bioinformatics tools that were designed, tested, and optimized for model species. Also, assembling the whole genome or transcriptome is only the first step. A critical point is having a good annotation, fundamental for any downstream analysis: many of the annotation methods used so far are based on sequence similarity. Since most of the protein or nucleotide sequences in databases come from model species, this method is more adequate for taxa that are reasonably close—from an evolutionary point of view—to such species. If this is not the case, a considerable part of genes will have either no annotation, or a low quality annotation; as a result, sequences labeled with “unknown protein” are very common in databases. This issue will be discussed more in detail in Chapter 2.

Another important point is the detection of orthologous genes. Orthologs are coding sequences that evolved from a common ancestral gene and diverged after a speciation event. Inferring orthology is a central point for comparative analyses, which, in turn, is fundamental for studying evolution. There are many methods for detecting orthologs but, even in this case, most of them are based on sequence similarity, therefore, when we compare species with phylogenetic distances of hundreds of millions years, only highly conserved orthologous genes will be identified, while those that are more variable, and likely more interesting for the study of evolution, will be discarded from the analyses (this point will be discussed in Chapter 3).

There is no simple solution for these problems about the study of non-model species, but an increase in available data and an implementation of bioinformatic tools to allow a proper comparison between distant taxa is necessary for a complete comprehension of many biological fields.










Species	Finding	
<i>Chelonia mydas</i> and <i>Pelodiscus sinensis</i>	Genome-wide phylogenetic analysis resolved the position of turtles within animals (Chiari et al. 2012).	
<i>Falco peregrinus</i> and <i>Falco cherrug</i>	Duplication of genes involved in beak morphology in response to predatory life style (Zhan et al. 2013).	
<i>Picea abies</i>	Accumulation of transposable elements is responsible for most of the 100 times larger genome size of spruce compared to other plant species (Nystedt et al. 2013).	
<i>Mnemiopsis leidyi</i> and <i>Pleurobrachia bachei</i>	Missing of HOX genes and classical neurotransmitter pathways provide insights into early evolution and neuronal system development (Moroz et al. 2014).	
<i>Latimeria chalumnae</i>	Gene loss associated with vertebrate transition from water to land (Amemiya et al. (2013).	
<i>Heliconius melpomene</i>	Gene duplication of opsin underlie visual complexity (Heliconius Genome Sequencing Consortium, 2012).	
<i>Crassostrea gigas</i>	Expansion of genes involved in protection against heat and stress (Zhang et al. 2012).	
<i>Petromyzon marinus</i>	Insights into genetic innovations, emerged at the base of vertebrates evolution (Smith et al. 2013).	
<i>Bos grunniens</i>	Expansion of gene families related to hypoxic stress in response to high altitudes (Qiu et al. 2012).	

Table 2. Example of key findings from genomes and transcriptomes of non-model species.

Why Mollusca?

With more than 100,000 extant species, Mollusca is the second Phylum for number of species after arthropods (Haszprunar et al. 2008). This taxon originated before Cambrian and contains eight extant classes: Bivalvia, Cephalopoda, Gastropoda, Scaphopoda, Polyplacophora, Monoplacophora, Caudofoveata, and Solenogastres. These Classes show a great diversity in term of morphology, size, complexity, and behavioral repertoires (figure 2). Molluscs are abundant in marine and terrestrial environments and some species can live in extreme conditions, such as hydrothermal vents. Some of species are invasive, like the Zebra mussel (*Dreissena polymorpha*), that causes every year millions of dollars of economic damage, while others are vectors of diseases and parasites. Furthermore, some bivalves (oysters, clams, mussels), gastropods (abalone, queen conch), and cephalopods (squid, octopus, and cuttlefish) represent a very important food source worldwide. Finally, other species are also important for their production of pearls (oysters) and shells. For all these reasons, molluscs are a fundamental component of fisheries and aquaculture.



Figure 2. An example of the great diversity within the phylum Mollusca. A) *Glaucus atlanticus* (Gastropoda); B) *Tonicella lineata* (Polyplacophora); C) *Ctenoides scaber* (Bivalvia); D) *Wirenia argentea* (Solenogastres); E) *Sepioteuthis sepioidea* (Cephalopoda); F) *Antalis vulgaris* (Scaphopoda).

Besides their economic and ecological importance, molluscs are a model for studying brain organization, learning, memory, and eye development (Haszprunar & Wanninger 2012; Ogura et al. 2004; Hochner & Glanzman 2016; Tamvacakis et al. 2015).

Considering their great diversity, the number of species, and their ancient history, molluscs have all the features to be good candidates for studying evolution, nevertheless, they are widely underrepresented for what concerns genomics data: Mollusca, among Metazoa, is the Phylum with the lowest ratio between fraction of sequenced genomes and number of described species (Schell et al. 2017; figure 3).

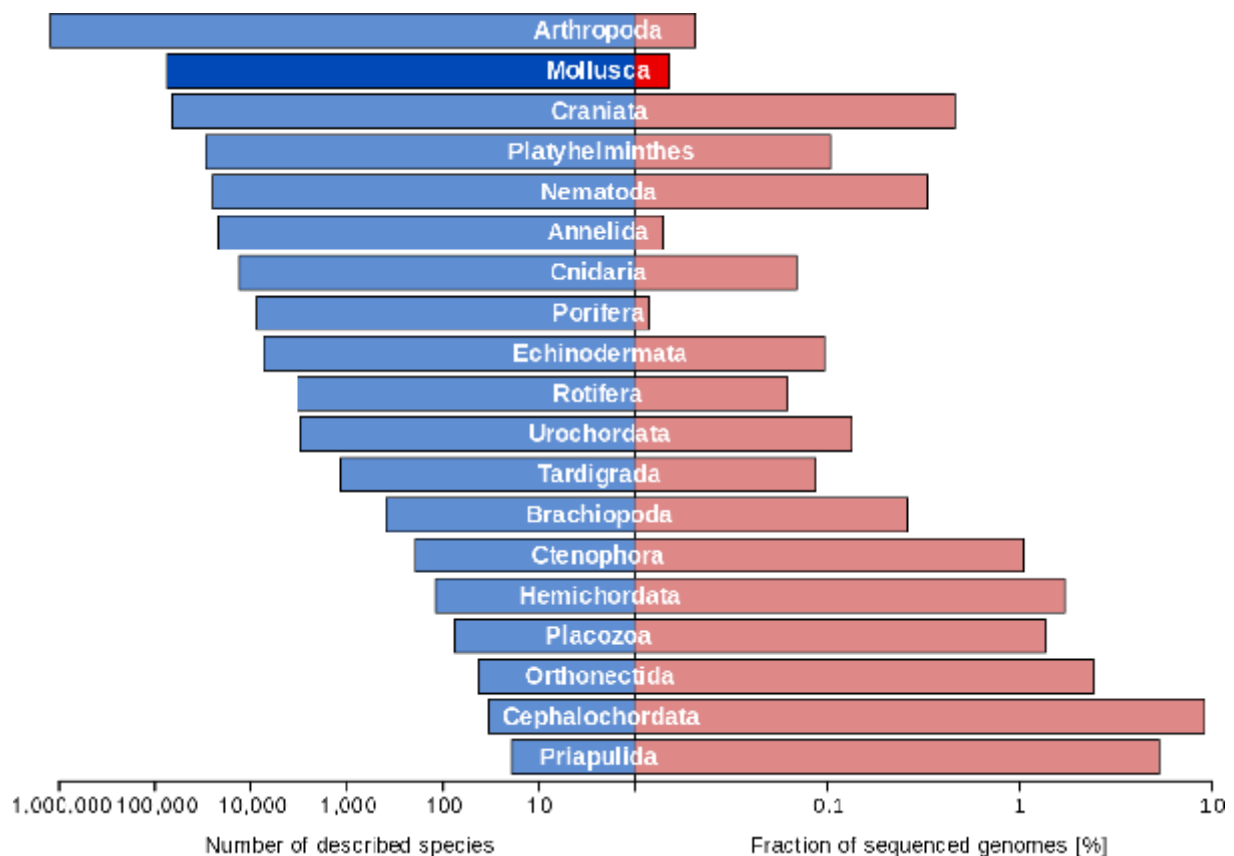


Figure 3. Number of described species and fraction of sequenced genomes in Animal phyla (Schell et al. 2017).

Indeed, only 12 genomes were published so far and they are not representative of the phylogenetic diversity, since five out of eight Classes have no reference

genomes. Also, the available mollusc genomes are very variable in terms of quality and completeness (table 3). The first to be assembled, in 2012, was the complete genome of the Pacific oyster *Crassostrea gigas* (Zhang et al. 2012), followed by the genomes of four other bivalves (*Pinctada fucata*, *Mytilus galloprovincialis*, *Patinopecten yessoensis* and *Dreissena polymorpha*), six gastropods (*Lottia gigantea*, *Patella vulgata*, *Aplysia californica*, *Conus tribblei*, *Radix auricularia* and *Biomphalaria glabrata*) and one cephalopod (*Octopus bimaculoides*) (Albertin et al. 2015). Although much more transcriptome data are available on public databases, few comparative transcriptomic works are present in the literature and most of the biological questions investigated so far remain unexplored in Mollusca.

If, on the one hand, nuclear genomes are far from being characterized in Mollusca, on the other hand, mitochondrial genomes (mtDNAs) have been more deeply investigated, mostly by using Sanger sequencing. Such studies revealed several peculiar features: compared to most Metazoa, mitochondrial genomes of molluscs can be particularly large (> 42 kb), have great variations in gene content—in terms of both gene gain and loss—and gene order, with common events of rearrangements (Breton et al. 2014). Also, mollusc mtDNAs are typically characterized by high substitution rates and a large fraction of non-coding regions (unassigned regions, URs). In addition, a unique mechanism of mitochondrial inheritance was found in ~100 bivalve species (Gusman et al. 2016) and it will be discussed in the next Chapter.

Species	Assembly length / Estimated	#sequences	N50	Technology	Coverage	BUSCO
	genome size = % assembled					
<i>Octopus bimaculoides</i>	2.4Gb/2.7Gb = 89%	151674	475 kb	Illumina	92	73.8
<i>Crassostrea gigas</i>	558 Mb / 890 Mb = 62.7%	7659	402 kb	Illumina	100	82
<i>Mytilus galloprovincialis</i>	1.6Gb / 1.9Gb = 86%	2315965	1067 bp	Illumina	17	1.6
<i>Pinctada fucata</i>	1,150 Mb / NA	29682	14,5 kb	454/ Illumina	40	NA
<i>Patinopecten yessoensis</i>	988 Mb / 1.43 Gb = 69%	145537	804 kB	Illumina	45	NA
<i>Dreissena polymorpha</i>	906Mb / 1.7 Gb = 53%	1057	855 bp	454	3	NA
<i>Lottia gigantea</i>	360 Mb / 421 Mb = 85%	4469	1870 kb	Sanger	8.87	97.0
<i>Patella vulgata</i>	579 Mb / 1,460 Mb = 39.7%	295348	3160 bp	Illumina	25.6	16.6
<i>Conus tribblei</i>	2,160 Mb / 2,757 Mb = 78%	1126156	2681 bp	Illumina	28.5	44
<i>Aplysia californica</i>	927 Mb / 1,760 Mb = 47%	4332	918 kb	Illumina	66	94.1
<i>Biomphalaria glabrata</i>	916 Mb / 929 Mb = 99%	331401	48 kb	454	27.5	89.1
<i>Radix auricularia</i>	910 Mb / 1.6 Gb = 57%	4823	578 kb	Illumina	72	94.6

Table 3. Genome sequencing statistics of molluscs species, including assembly length, estimated genome size, % of assembled genome, n° of sequences obtained from the assembly , N50 length, technology used for the sequencing, coverage and a measure for the assessment of genome assembly completeness provided by BUSCO (Simão et al. 2015) .

An unusual mechanism of mitochondrial inheritance

Metazoans are characterized by a strictly maternal inheritance of mitochondria (SMI), namely only females transmit mitochondria to the offspring, while paternal mitochondrial contribution is avoided in very different ways across eukaryotes (Sato & Sato 2013). It is commonly accepted that SMI appeared independently many times in the evolution of Eukaryotes, suggesting a strong evolutionary advantage of uniparental inheritance, such as limiting the spread of endosymbiotic parasites, lethal mutations and selfish genetic elements, and reducing intergenomic conflicts (Birky 1995). Until now, the only known evolutionarily stable exception to SMI in Metazoa is the Doubly Uniparental Inheritance (DUI) (see Zouros 2013 for a detailed review). So far, DUI has been reported in ~100 species of seven families of bivalve molluscs (Donacidae, Hyriidae, Margaritiferidae, Mytilidae, Solenoidae, Unionidae and Veneridae; Gusman et al. 2016). In DUI species two distinct mitochondrial lineages are present: the F-type, inherited through eggs, and the M-type, inherited through sperm. The embryo, upon fertilization, receives both types of mitochondria, but if it develops into a female, the M-type gets degraded or diluted, and the adult female will be homoplasmic for the F-type. On the contrary, if the embryo develops into a male, the M-type will be predominant in the gonad, and present in variable amounts in the somatic tissues (Ghiselli et al. 2011). Consequently, the adult male will be heteroplasmic, but his gametes will be homoplasmic for the M-type (figure 4).

Most of the issues about origin, evolutionary advantages (if any), as well as molecular mechanisms of DUI are still unknown. Figure 5 shows the distribution of DUI species within the Bivalvia class. The presence of this peculiar mechanism of mitochondrial inheritance is scattered across the phylogenetic tree, with entire families characterized by SMI. On the contrary, within families where DUI was detected, some species are characterized by SMI, while others by DUI. This evidence opens questions about the origin of DUI; if DUI appeared once during the evolution of Bivalvia, its origin should be dated in concomitance to the evolutionary

radiation of Autolamellibranchia (at the beginning of Ordovician), and this peculiar inheritance mechanism should have been subsequently lost in many taxa (Zouros 2013). Otherwise, DUI could have been originated many times during the evolution of Bivalvia; in this case, multiple and independent origins might imply some sort of evolutionary advantage and/or functional role of this mechanism. In line with the hypothesis of a multiple emergence of DUI, a viral origin was proposed by (Milani et al. 2016): according to this theory, a virus could have infected some mitochondria, providing the organelles with the ability to avoid degradation in embryo and invade the germ line.

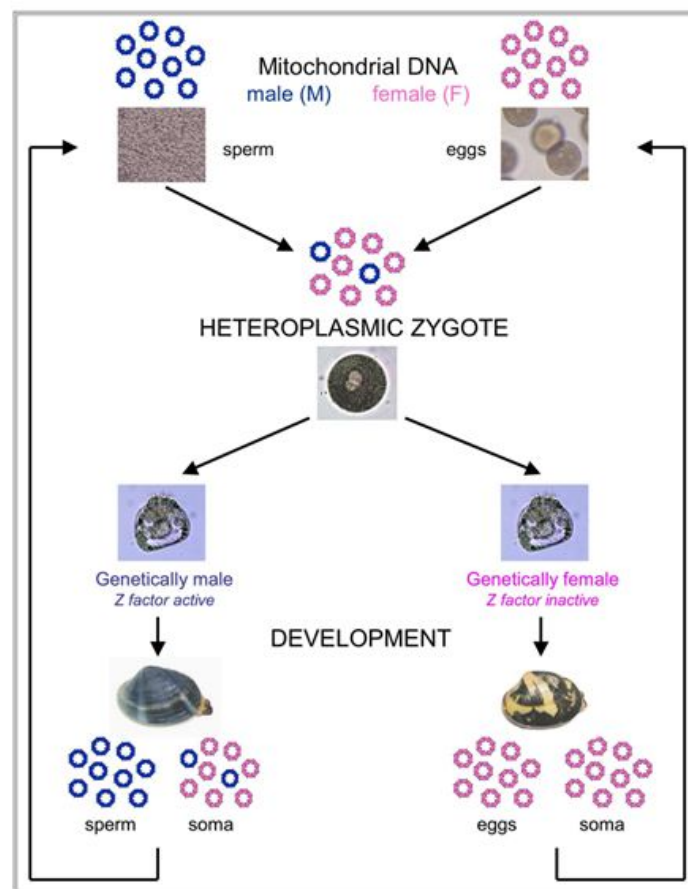


Figure 4. Schematic illustration of DUI mechanism (Passamonti & Ghiselli 2009).

The peculiar feature of DUI species is therefore the natural heteroplasmy of mtDNA in males. M-type and F-type genomes are very different in sequence: the amino acid p-distance varies according to the species, with lowest values in mytilids (~14%), highest in unionids (up to 52%) and with intermediate values in venerids (~34%)

(Zouros 2013; Doucet-Beauprè et al. 2010). Also, differences in gene content and gene order are common between the two mtDNAs. Particularly, a second copy of the gene *cox2*—120 bp longer than the original gene—was found in the M genome of *Musculista senhousia* (Passamonti et al. 2011). Similarly, the M-type genomes of Unionids show an extension of 48-192 codons at the 3' end of the *cox2* gene (Breton et al. 2009). Interestingly, a duplication of *cox2* was detected also in *Ruditapes philippinarum*, but in the F-type genome.

Finally, mitochondrial genomes of DUI species often harbor open reading frames with unknown homology and function (ORFans), located in the Unassigned Regions (URs) typical of most bivalve mtDNA (see Milani et al. 2013). ORFans were found in both M-type and F-type genomes of several DUI species and they are characterized by different lengths and positions. Although ORFans were initially considered as non-coding, the protein product of male and female ORFans was detected in gonads of *Venustaconcha ellipsiformis*, and a sex-specific function of such proteins was proposed by Breton et al. (2009). In addition, transcription of a M-type ORFan (*rphm21*) in *R. philippinarum* was reported in gonads, adductor and mantle of males (Ghiselli et al. 2013; Milani, Ghiselli, Iannello, et al. 2014). The protein RPHM21 was detected by Western blot and confocal microscopy in gonads and early embryos, as well (Milani, Ghiselli, Maurizii, et al. 2014).

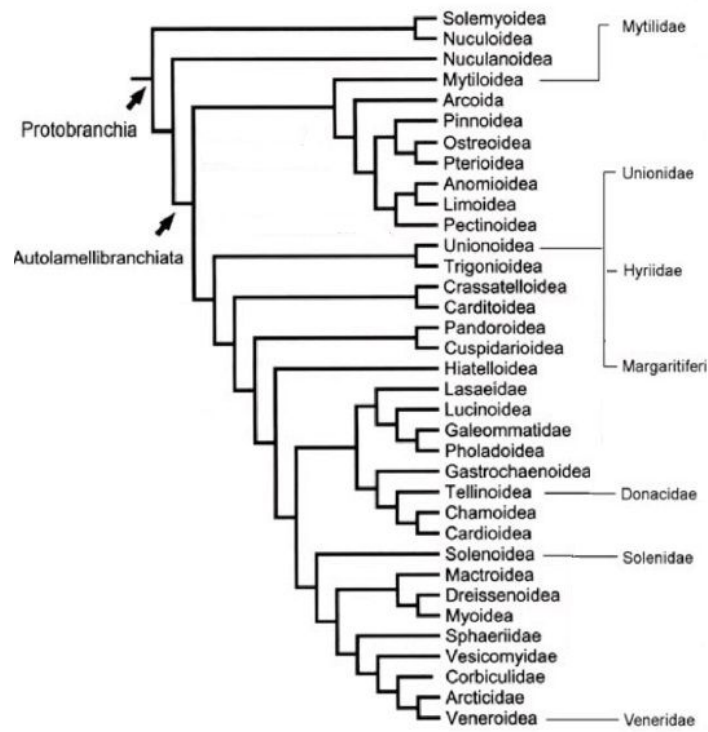


Figure 5. Phylogenetic distribution of DUI families within Bivalvia class.

Aim of this thesis

My research activity as PhD student was twofold: one, more general, is to start creating a comparative framework to analyze HTS data in non-model species (bivalve molluscs in particular); the other is to use such methodological structure to study the evolution of two bivalve species—*Ruditapes decussatus* and *Ruditapes philippinarum* (Bivalvia, Veneridae; figure 6)—focusing on sex biased genes and mito-nuclear coevolution.



Figure 6. *Ruditapes decussatus* (left) and *Ruditapes philippinarum* (right).

While *R. decussatus*—also known as grooved carpet shell—is native to the Mediterranean and European Atlantic coasts, *R. philippinarum*—also known as Manila clam—is native from Philippines, Korea, and Japan. Although the fishing of *R. decussatus* had historically a main role in the production of seafood in Italy, Spain and Portugal, the recent introduction in Europe of *R. philippinarum* led to a replacement of *R. decussatus* with *R. philippinarum* for aquaculture purposes, and to a consequent decline of *R. decussatus* population in the Southwestern Europe. In fact, compared to *R. decussatus*, *R. philippinarum* is preferred in fisheries, since it reaches sexual maturation at a smaller size, is faster growing, has a greater number of spawning events per year, a more extended breeding period, and a higher resistance to disease (Gosling 2003).

Like some other bivalve species, *R. philippinarum* is characterized by DUI (Passamonti & Scali 2001) (see the paragraph “An unusual mechanism of mitochondrial inheritance”). Before the work in this thesis was carried out, no

information was available about the kind of mitochondrial inheritance of *R. decussatus*. The results of my first analyses showed that *R. decussatus* is a SMI species (Chapter 2 and Ghiselli et al. 2017), so it was used, in the following works, to implement comparative analyses between congeneric species having a different mitochondrial inheritance mechanism (Chapters 3 and 4). Moreover, during my PhD, I developed a pipeline specifically designed for the annotation of transcriptomes in non-model animals (Chapter 1), then I used RNA-Seq data to: 1) obtain the mitochondrial genome of *R. decussatus*, validate it by Sanger sequencing, and investigate the kind of mitochondrial inheritance in this species (i.e.: DUI vs SMI; see Chapter 2 and Ghiselli et al. 2017); 2) investigate transcription data from mature gonads of *R. decussatus* and *R. philippinarum*, focusing on the evolution of sex-biased genes (Chapter 3); 3) analyze sequence and transcription variation in the subunits of the oxidative phosphorylation complexes between the two species, focusing on mito-nuclear coevolution (Chapter 4). I structured my work in the four following chapters, which correspond to 4 papers: one has been published in PeerJ (Chapter 2), one has been submitted to Genome Biology and Evolution (pending major revision Chapter 3), and two in preparation (Chapters 1 and 4):

- 1) A transcriptome annotation pipeline for non-model organisms;
- 2) The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae);
- 3) Comparative transcriptomics in two bivalve species offers different perspectives on the evolution of sex-biased genes;
- 4) No evidence for nuclear compensation hypothesis in species with different mechanisms of mitochondrial inheritance.

Short summaries of the 4 articles are reported in the next paragraph.

Short summaries

1. A transcriptome annotation pipeline for non-model organisms

The introduction of high-throughput sequencing technologies allowed researchers to generate large amounts of genomic data at limited cost and time. This opportunity had a groundbreaking impact on the study of non-model organisms: above all, RNA-Seq and *de novo* transcriptome assembly represent a valuable source of information in species for which genomic resources are scarce or absent. However, sequencing and assembly are only the first steps, and an accurate annotation is fundamental for every kind of biological analysis. Annotation of transcriptomes from model organisms and their closely-related species is quite straightforward, and is generally based on simple sequence similarity searches. Conversely, non-model organisms require more complex and integrated procedures in order to infer remote homology and function. I present a pipeline specifically thought for the annotation of transcriptomes of non-model organisms. It consists of an integrated approach that combines different bioinformatics tools to obtain: 1) filtration from contaminant sequences; 2) ORF prediction, identification of pseudogenes and artificially fused transcripts; 3) annotation at the amino acid level, based both on sequence similarity and on the identification of conserved domains by protein signature recognition, functional annotation by the assignment of GO terms; 4) identification of orthologs and paralogs; 5) annotation at the nucleotide level; 6) low quality annotation and protein feature prediction. I tested this pipeline by re-annotating the transcriptome of *Ruditapes philippinarum* (Bivalvia, Veneridae).

2. The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae)

Molluscs, and particularly bivalves, show a great diversity of mitochondrial genomes, in terms of size (up to ~47Kb) and features that are not common in most of Metazoa, such as high proportion of non-coding sequences (URs), gene rearrangements, differences in strand usage and novel protein coding genes with unknown function

(ORFans). Also, a peculiar feature that characterized mtDNA of around 100 bivalve species is the doubly uniparental inheritance (DUI) (see the paragraph “An unusual mechanism of mitochondrial inheritance”). The detection of DUI is not a straightforward process when we sequence mitochondrial genomes, especially using PCR-based approaches: in fact, since M-type and F-type can exhibit a nucleotide p-distance above 40%, primers may fail to amplify one of the two mtDNAs, yielding a false-negative result. Although, HTS approach has been scarcely used in analyzing mitochondrial genome and transcriptome data, it can easily detect low copy mitochondrial variants.

In this work, I used RNA-Seq data to assemble, for the first time, the complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus*, revealing the presence of a unique mtDNA variant in both males and females and, therefore, a strictly maternal inheritance of mitochondria in this species. The mitochondrial genome was also validated by Sanger sequencing. Also, I performed an analysis on sequence polymorphism (SP) and structural variants among the samples. Finally, I compared the mitochondrial features of *R. decussatus* with those of other venerid bivalves.

3. Comparative transcriptomics in two bivalve species offers different perspectives on the evolution of sex-biased genes

Despite the differences in terms of sexual dimorphism and behavior, males and females share almost the same genome. Therefore, the vast majority of sex-specific characters and traits are the result of differential expression of the so-called ‘sex-biased genes’. Previous works reported that, in many organisms, genes that are more or exclusively transcribed in males (male-biased genes) show a higher rate of protein evolution—calculated as the ratio of nonsynonymous to synonymous nucleotide substitution (dN/dS). Particularly, male-biased genes seem to be the most divergent also in terms of transcription level. This evidence inspired the hypothesis of a positive correlation between the evolution of protein sequences and transcriptional divergence. Although the study of sex-biased genes is crucial for understanding the mechanisms of gene regulation and evolution, investigations on evolution and

transcription of such genes are missing for many taxa and no data about this topic are available on Mollusca.

Another central point for understanding molecular evolution is what shapes the rate of protein sequence change. According to the most recent theories, transcription level has been proposed to be the main responsible for the rate of protein evolution a strong negative correlation was found between dN/dS and transcription level, across the three domains of life.

In this work I obtained the gonadal transcriptome from males and females of the European clam *R. decussatus* and I compared it with the available gonadal transcriptome data (Ghiselli et al. 2012) from the species *R. philippinarum*. This comparison allowed to investigate, for the first time in two bivalve species, the evolution of both protein sequence and transcription level divergence of sex-biased genes. I also characterized the biological processes represented in male and female mature gonads of the two species. Finally, I investigated the relationship between the transcription level and rate of protein evolution in both *R. decussatus* and *R. philippinarum*.

4. No evidence for nuclear compensation hypothesis in species with different mechanisms of mitochondrial inheritance

Mitochondria are a fundamental component of the eukaryotic cell. Although they are involved in many biological processes, a central role of mitochondria is the production of ATP through oxidative phosphorylation (OXPHOS). Genes involved in this metabolic pathway are encoded by both nuclear and mitochondrial genome and a tight co-evolution and co-regulation of these two genomes is essential to maintain efficient mitochondrial activity. Indeed, many works investigated the effects of having different mtDNA variants working with the same nuclear background by producing cytoplasmic hybrids. These works shows detrimental effects of heteroplasmy, such as reduction of OXPHOS activity, oxidative damage, disruption of mitochondrial functions. All these evidences supports the theory that Metazoa have evolved a non-mendelian mechanism of mitochondrial inheritance in order to avoid the

presence of mixed mtDNA haplotypes in the same organism (Lane 2012), reducing the emergence of genomic conflicts and mito-nuclear incompatibilities.

Although generating cytoplasmic hybrids is the most common method used to investigate the effects of heteroplasmy, and thus the dynamics of mito-nuclear co-evolution, the heteroplasmic condition is artificial and the biological processes may be affected by this. On the contrary, in the DUI male heteroplasmy is natural, therefore its biological functions and interactions between nucleus and mitochondria are the unaltered result of evolution. In this work I took advantage of the natural heteroplasmic condition of DUI species to get insights into the dynamics of mito-nuclear co-evolution. I compared the transcriptomes of the SMI species *R. decussatus* and the DUI species *R. philippinarum*. I investigated the rate of protein evolution and the transcription of both nuclear and mitochondrial genes involved in mitochondrial functions and I particularly focused on genes encoding for subunits of OXPHOS complexes.

Chapter 1:

A transcriptome annotation pipeline for non-model organisms

Mariangela Iannello¹, Guglielmo Puccio, Fabrizio Ghiselli^{*.1.°}, Marco Passamonti^{1.°}

¹Department of Biological, Geological and Environmental Sciences – University of Bologna, Italy.

*Corresponding Author

°Equal contribution

Introduction

The development of High Throughput Sequencing (HTS) had a profound impact on the study of non-model species. In particular, RNA-Seq data allow scientists to investigate biological issues in species where a genome reference is not available, and with reduced times and affordable costs. Nevertheless, a good quality transcriptome annotation is not easy to achieve, and most of the bioinformatics tools are usually optimized for model species and organisms for which there is abundant information in the databases. Therefore, a multi-software, integrated approach is necessary in order to infer remote homology and predict function in non-model species. Also, being unfamiliar with the Unix environment, command line tools, and scripting—necessary for handling and analyzing large datasets such those produced by HTS—may represent an obstacle for many researchers. For this reason, we developed a transcription annotation pipeline specifically for non-model organisms. It combines different tools and databases in order to filter the results from contaminants and maximize the detection of remote homology. This pipeline is composed by six main steps, aiming to obtain: 1) filtering from contaminant sequences; 2) ORF prediction, identification of pseudogenes and artificially fused transcripts; 3) amino acid sequence annotation, based on sequence similarity and on the identification of conserved domains by protein signature recognition, and functional annotation by the assignment of GO terms; 4) identification of orthologs and paralogs; 5) nucleotide sequence annotation (for noncoding transcripts); 6) low quality annotation and protein feature prediction (figure 7).

Each step of the pipeline is described separately in this chapter, and all the scripts used are reported in an appendix at the end of each module. Most of the tools used are stand-alone (i.e. they can be installed locally), except for Argot2 (step 3) and OrthoMCL (step 4) that requires access to the web. The Bash language was used for most of the scripts and an automated version in Python is almost ready to make this pipeline more “friendly” for researches that have no experience in shell scripting.

Transcriptomes typically consist of tens of thousands sequences, and the annotation of such a large amount of data requires high performance computing. Softwares like BLAST, HMMER and InterProScan work faster using multiple CPUs: in that way the

programs can assign a process to each available thread and run them in parallel, saving time. For this reason, the use of a high performance computing cluster, or at least of a powerful multicore workstation is strongly recommended. If not possible, a good alternative is cloud computing, like for example Amazon EC2. It is also important to take into account that the softwares and the databases used in the pipeline require an appropriate storage space.

Performing searches on tens of thousands sequences is computationally intensive and can take several days. A possible solution is to split the input fasta file into n parts—depending on the number of available nodes—and run each part on a node/CPU (the so-called “*divide et impera*” approach). In this way it is possible to get parallelization from tools that are normally serial. Of course, each process can be run on multiple threads in a node.

To test this pipeline, we re-annotated a *R. philippinarum* gonadal transcriptome previously annotated using BLAST2GO and BLASTX in (Ghiselli et al. 2012). Compared to the previous annotation, this pipeline yielded 15% more annotated genes. General statistics for each annotation step are reported in table 4.

A detailed step-by-step description of the pipeline, as well as all the data of *R. philippinarum* annotation, are available online as an Open Science Framework (OSF, www.osf.io) project:

https://osf.io/cdkb9/?view_only=f0b2cde926db43719f3d705012c4eeaa

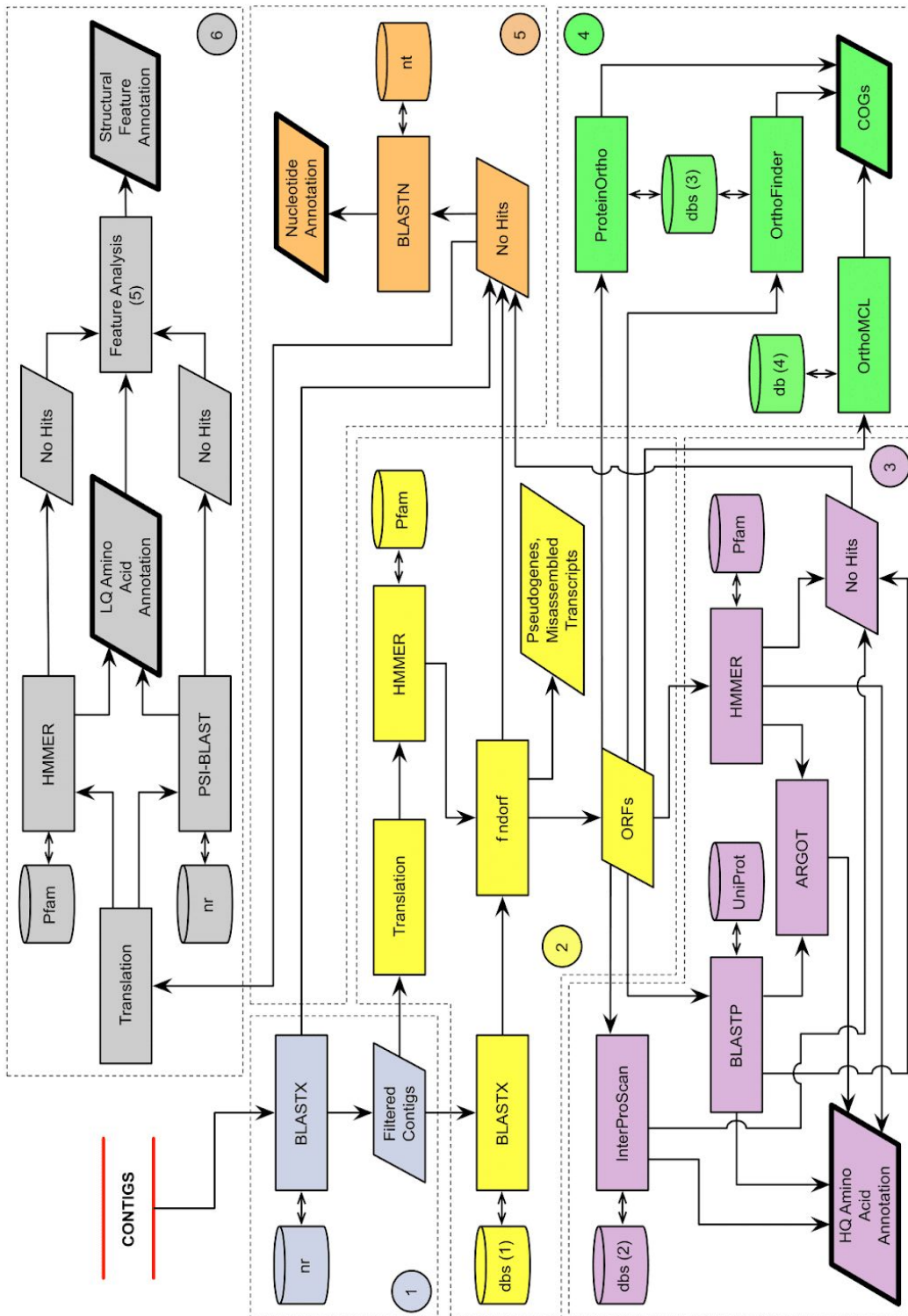


Figure 7. Transcription annotation pipeline for non-model organisms. Different colors represent different annotation steps. An example of databases (1,2,3,4) and software (5) used for the annotation of *R. philippinarum* is reported in table 5.

1) Contaminant Filtering

The first module of the annotation pipeline aims at removing all the contaminant sequences, namely transcripts not coming from the organism(s) of interest for the study. The problem of contaminants is well known in HTS data, nevertheless a filtering step is often omitted. Contaminants can be of different types and origin, depending on the organism (e.g. ecology, behavior), and the sampled tissue; common sources of contamination are the environment (e.g. air, soil, water), symbionts and/or parasites, food, etc. This step is extremely important to avoid data misinterpretation and the attribution of sequences to the wrong organism. The second mistake is particularly detrimental because it propagates through the databases yielding erroneous and misleading search results (Merchant et al. 2014).

In principle, filtering can be performed before the assembly, directly on the reads, or after the assembly, on the contigs. The former approach has the advantage that contaminant reads are excluded from the downstream processes, reducing the possibility of generating chimeric contigs, and speeding up the assembly. Such kind of approach (e.g.: blobtools, formerly blobology, see Kumar et al. 2013; Laetsch & Blaxter 2017) was successfully applied to identify and remove an extensive contamination from the genome of the tardigrade *Hypsibius dujardini* (Koutsovoulos et al. 2016). In that case, before it was identified, the contamination led to a wrong interpretation of the data (i.e.: massive horizontal gene transfer, see Boothby et al. 2015). That said, this approach cannot be used on transcriptomes, so a “post-assembly” approach must be taken.

In this pipeline, the removal of contaminant contigs is achieved through a BLASTX search (Altschul et al. 1990) performed on the assembly against the nr database. By using proper flags, BLASTX will output the first 10 best hits for each query, and the taxon ID of the subject (staxids) for each matching organism (see appendix A for details). Next, we use the “genomes” Bioconductor library to obtain the full taxonomic lineage from the BLAST staxids. Then, by using the custom script “filter.sh”, each staxids will be associated to the full taxonomic lineage and so that the filtering is

possible at any taxonomic level, by keeping only contigs that have a strong match with the chosen taxon.

In the annotation of *R. philippinarum*, we decided to keep only the queries that did have a strong match with 'Metazoa'. This step allowed the removal of 1.7% of the contigs, coming from non-Metazoa contaminants (table 4).

Appendix A

Blastx

```
blastx -db <nr> -query Rph_contigs_unwrapped.fasta -evaluate 1e-10 -  
max_target_seqs 10 -outfmt "6 qseqid sseqid staxids evaluate" -num_threads  
<n> -out Rph.blastx_1
```

```
Select 'staxids'
```

```
awk -F $'\t' '{print $3}' Rph.blastx_1 | sed 's;/\n/g' | sort -u >  
Rph_blastx_1.staxids
```

Assign taxonomy lineage

```
source("https://bioconductor.org/biocLite.R")
biocLite("genomes")
library(genomes)

Sys.setenv(email='username@host.domain')

staxids = read.table("Rph_blastx_1.staxids", header=FALSE)
i <- 1
while (i < length(staxids$V1)) {
  ids=staxids$V1[i]
  fulltaxa=ncbiTaxonomy(ids, summary=FALSE)
  write.table(fulltaxa, "lineages.txt", append=TRUE,
    col.names=FALSE, row.names=FALSE, sep="\t",
    quote=FALSE)
  i <- i + 1
}
```

Filter with filter.sh scripps

```
#!/bin/bash

dotpid=
rundots() { ( trap 'exit 0' SIGUSR1; while : ; do echo -n '.' >&2; sleep
0.2; done) & dotpid=$!; }
stopdots() { kill -USR1 $dotpid; wait $dotpid; trap EXIT; }
startdots() { rundots; trap "stopdots" EXIT; return 0; }

echo "Input taxon name:"
read taxon
echo 'Please Wait'
startdots
grep -w $taxon lineages.txt | awk '{print $1}' | sort -u > $taxon.staxids
awk '{print $1"\t"$2"\t"$3"\t"$4}' Rph.blastx_1 > Rph_blastx_1.fixed
file=$taxon.staxids ; name=$(cat $file) ; for ids in $name ; do grep
";"$ids";" Rph.blastx_1.fixed ; done > Rph_"$taxon".blastx_1
cut -f1 Rph_"$taxon".blastx_1 | sort -u > Rph_"$taxon"_blastx_1.ids
stopdots
echo
echo "The number of transcripts matching $taxon is:" $(cat
Rph_"$taxon"_blastx_1.ids | wc -l)
echo
echo "Generating FASTA file of sequences matching $taxon"
echo "This might take a while, please wait"
startdots
file=Rph_"$taxon"_blastx_1.ids ; name=$(cat $file); for locus in $name; do
grep -w -A1 $locus ../Rph_contigs_unwrapped.fasta; done >
Rph_"$taxon"_blastx_1.fasta
stopdots
echo
echo "Done."
```

2) ORF prediction

Detection of ORFs is an essential step in the annotation of a whole transcriptome. In principle, any sequence between a start and a stop codon in each assembled contig could be a coding sequence. However, each contig can include tens of putative ORFs but, only one—or at most a few of them—will code for a protein. Therefore, a program for the prediction of coding ORFs is essential to filter non-coding sequences and make computational work less intense. In this module, we use ‘findorf’ for this purpose. Findorf is an ORF prediction tool— developed by Vincent Buffalo and tested in Krasileva et al. (2013)—specifically for non-model organisms (<https://github.com/vsbuffalo/findorf>). Compared to other ORF prediction tools, this includes a BLASTX search against separate user-defined databases from closely-related species. The annotation reported in databases from non-model species are almost always predicted automatically without human supervision, and may include errors. The advantage of this approach is that, by using different databases, each search will be validated by multiple independent sources. An example of databases used in the ORF prediction of *R. philippinarum* is shown in table 5. Furthermore, this tool performs an ‘hmmscan’ search—as implemented in HMMER (Finn et al. 2011)—against the Pfam database (Finn et al. 2014) in order to predict ORFs using Hidden Markov Model (HMM) profiles. Since HMMER needs amino acid sequences as input file, we use the EMBOSS ‘transeq’ tool (Rice et al. 2000) to translate contigs in all the six frames. Finally, findorf identifies also frameshift mutations, premature stop codons, misassembled transcripts, and pseudogenes.

Appendix B

Blastx

```
blastx -db <database_name> -query Rph_Metazoa_blastx.fasta -evaluate 1e-3 -  
outfmt 5 -num_threads <n> -out <xml_output.blastx>
```

Translate contings in the 6 frames

```
transeq -sequence Rph_Metazoa_blastx_1.fasta -outseq  
Rph_Metazoa_allframes.fasta -alternative -frame 6 -trim T
```

Fix header

```
sed -e 's/_1$/_+1/;s/_1 /_+1 /' -e 's/_2$/_+2/;s/_2 /_+2 /' -e  
's/_3$/_+3/;s/_3 /_+3 /' -e 's/_4$/_-1/;s/_4 /_-1 /' -e 's/_5$/_-2/;s/_5  
/_-2 /' -e 's/_6$/_-3/;s/_6 /_-3 /' Rph_Metazoa_allframes.fasta >  
Rph_Metazoa_allframes_fixed.fasta
```

HMMER

```
hmmsearch -E 0.001 --domE 1 --tblout Rph_Metazoa_hmmer_2.tblout --domtblout  
Rph_Metazoa_hmmer_2.domtblout -o Rph_Metazoa_hmmer_2.out --noali --cpu <n>  
<Pfam_database.hmm> Rph_Metazoa_allframes_fixed.fasta
```

Fix HMMER domtblout

```
python hmmerfix.py Rph_Metazoa_hmmer_2.domtblout >  
Rph_Metazoa_hmmer_2.domtblout.tab
```

Findorf join

```
findorf join --ref Rph_Metazoa_blastx.fasta [blastx results aginats each  
database]--domain-hits Rph_Metazoa_hmmer_2.domtblout.tab
```

Finforf predict

```
findorf predict --input joined_blastx_dbs.pkl --evaluate 1e-5 --verbose \  
--gtf Rph_Metazoa_orf.gtf \  
--orf Rph_Metazoa_orf.fasta \  
--protein Rph_Metazoa_proteins.fasta \  
--frameshift Rph_Metazoa_frameshift.fasta \  
--stop Rph_Metazoa_stop.fasta \  
--no-relatives Rph_Metazoa_orfans.fasta \  
--masked Rph_Metazoa_masked.fasta --use-pfam
```

3) High-quality amino acid sequence annotation

The central module of this pipeline consists in a high-quality amino acid sequence annotation. For this purpose, we use different approaches to annotate the ORFs predicted in step 2, and transcripts are annotated by using both sequence similarity search and identification of protein domains.

A sequence similarity search is obtained with BLASTP (Altschul et al. 1990) and HMMER. we use BLASTP against the UniProt database (The UniProt Consortium, 2015): in this way a sequence similarity search is performed against a high quality protein sequence database. On the other hand, a search based on Hidden Markov Model (HMM) profiles is performed by HMMER against the Pfam database. BLASTP and HMMER outputs are combined together by using Argot2 (Falda et al. 2012). This tool allows to provide a function prediction by assigning Gene Ontology (GO) terms (Ashburner et al. 2000). Since a stand-alone version of Argot2 is not available, this step needs the BLASTP and HMMER outputs to be uploaded on the Argot2 web server (<http://www.medcomp.medicina.unipd.it/Argot2/>) and the Argot2 results to be downloaded. This time consuming step has been automatized in our annotation pipeline, so that it can be executed from command line.

In addition to the sequence similarity searches, we use InterProScan (Jones et al. 2014) to protein domains. InterProScan is a software developed to identify predictive signatures, provided by multiple databases. An example of signature databases, used in the annotation of *R. philippinarum*, is shown in table 5. InterProScan provides also sequence function prediction by the assignment of GO terms. Figure 8 shows the contribution of each tool in the annotation of *R. philippinarum* ORFs.

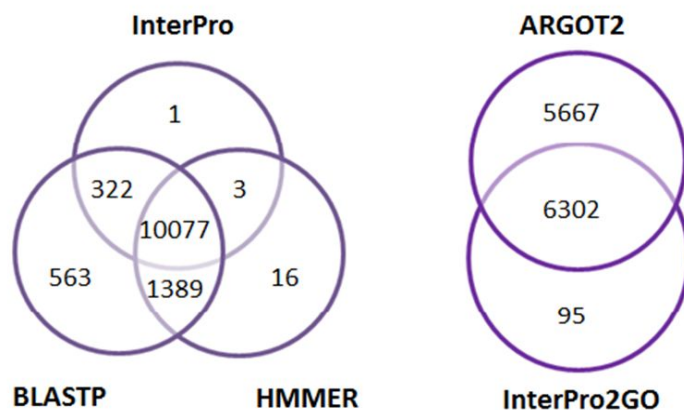


Figure 8. Contribution of InterProScan, BLASTP and HMMER in the high quality amino acid sequence annotation (left) and the contribution of Argot2 and InterPro2GO in the functional annotation (right)

Appendix C

Blastp

```
blastp -query Rph_Metazoa_proteins.fasta -db uniprot -outfmt "6 qseqid
sseqid evalue " -num_threads <n> -out Rph_proteins.blastp
```

HMMER

```
hmmsearch --tblout Rph_proteins.tblout --domtblout Rph_proteins.domtblout -o
Rph_proteins.o <PfamAB.hmm> Rph_Metazoa_proteins.fasta
```

InterProScan

```
interproscan.sh -appl TIGRFAM-13.0,Panther-8.1,SMART-6.2,PrositePatterns-
20.97,SuperFamily-1.75,PRINTS-42.0,Gene3d-3.5.0,PIRSF-
2.84,PrositeProfiles-20.97 -dp -f tsv -goterms -i
Rph_Metazoa_proteins.fasta -iprlookup -o Rph_proteins.interpro -T ./tmp
```

4) Identification of orthologs and paralogs

Detection of orthologs and paralogs is a crucial step for any comparative analysis. As for many analyses, the identification of orthologs with the available tools is optimal in model species (or species that are closely-related to them), therefore a combination of different tools is essential in this case to find clusters of orthologous groups (COGs).

OrthoMCL (Li et al. 2003) is one of the most widely used softwares for orthology prediction. This program clusters the input sequences into orthologous groups using a reciprocal best BLAST hit approach (Fischer et al. 2011). OrthoMCL interrogates the OrthoMCL-DB (Chen et al. 2006) public database that contains orthologous groups pre-computed using the OrthoMCL algorithm. Nevertheless, non-model organisms are often underrepresented—or missing completely—in OrthoMCL-DB; for example, it does not include any sequence from Spiralia. For this reason, we also use ProteinOrtho (Lechner et al. 2011) and Orthofinder (Emms & Kelly 2015), two stand-alone programs that implement an extended version of the reciprocal best hit method, but, most importantly, that allow to create custom databases and therefore to search for orthologs in species that are closely-related to the one investigated. Table 5 shows a list of species that we used to create the database used with ProteinOrtho and Orthofinder. Figure 9 shows the contribution of each tool in the detection of COGs.

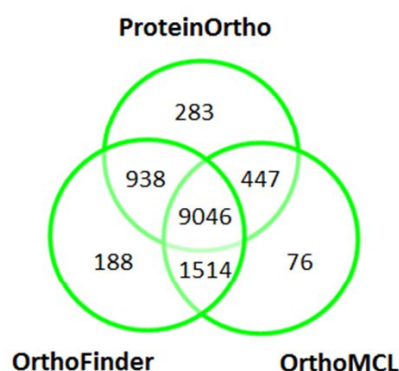


Figure 9. Contribution of ProteinOrtho, Orthofinder and OrthoMCL in the detection of clusters of orthologous groups

Appendix D

ProteinOrtho

```
perl proteinortho5.pl -project=mollusca_annelida -clean -verbose \  
-singles ../peptide_dbs_fasta/Cgi_peptides.fasta \  
../peptide_dbs_fasta/Lgi_peptides.fasta \  
../peptide_dbs_fasta/Obi_peptides.fasta \  
../peptide_dbs_fasta/Pfu_peptides.fasta \  
../peptide_dbs_fasta/Rde_metazoa_proteins.fasta \  
../peptide_dbs_fasta/Rph_metazoa_proteins.fasta \  
../peptide_dbs_fasta/Cte_peptides.fasta \  
../peptide_dbs_fasta/Hro_peptides.fasta
```

OrthoFinder

```
for files in *.fasta; do echo $files; grep -c ">" $files; awk '{print $1}'  
$files | grep ">" | sort -u | wc -l; done
```

```
python orthofinder.py -f ../peptide_dbs_fasta -t <n>
```


5) Nucleotide annotation

A nucleotide sequence annotation is performed in this step by using BLASTN (Altschul et al. 1990) against the nt database. This module was thought to annotate sequences that were not included in the previously steps, such as:

1. contigs that did not get hits from the BLASTX search in step 1;
2. contigs for which no ORFs were predicted in step 2;
3. contigs with no amino acid sequence annotation in step 3.

This search allows to infer ORFs that for some reason were not included in the previously analyses as well as to identify non-coding sequences that may be found in RNA-Seq data, irrespective of the kind of library preparation used.

Appendix E

Noblastx contigs

```
grep ">" Rph_contigs_unwrapped.fasta | sed 's/\>/g' | sort -u >
Rph_contigs.ids
cut -f1 ../1_filtering/Rph.blastx_1 | sort -u > Rph_blastx.ids
comm -23 Rph_contigs.ids Rph_blastx.ids > Rph_noblastx.ids
```

Noorf contigs

```
grep ">" Rph_Metazoa_blastx.fasta | sed 's/\>///g' | sort -u >
Rph_Metazoa.ids
grep ">" Rph_Metazoa_orf.fasta | cut -d ' ' -f1 | sed 's/\>///g' | sort -u
> Rph_Metazoa_orf.ids
comm -23 Rph_Metazoa.ids Rph_Metazoa_orf.ids > Rph_Metazoa_noorf.ids
```

Nohq contigs

```
cut -f1 Rph_proteins.blastp | sort | uniq > Rph_blastp.ids
grep "Locus" Rph_proteins.tblout | awk '{print $3}' | sort | uniq >
Rph_hmmer_3.ids
cut -f1 Rph_proteins.interpro | sort | uniq > Rph_interpro.ids
comm -23 Rph_Metazoa_orf.ids Rph_blastp.ids > noblastp.tmp
comm -23 noblastp.tmp Rph_hmmer_3.ids > noblastpnohmmer.tmp
comm -23 noblastpnohmmer.tmp Rph_interpro.ids | sort > Rph_nohq.ids
rm noblastp*.tmp
```

Merging Noblastx, Noorf and Nohq contigs

```
cat Rph_noblastx.ids Rph_Metazoa_noorf.ids Rph_nohq.ids | sort >
Rph_nohits123.ids
file="Rph_nohits123.ids"; name=$(cat $file); for locus in $name; do grep -
w -A1 $locus Rph_contigs_unwrapped.fasta; done > Rph_nohits123.fasta
```

Blastn

```
blastn -task blastn -query Rph_nohits123.fasta -db nt -evalue 0.01 -outfmt
"6 qseqid sseqid qstart qend sstart send stitle staxids length pident
evaluate " cd -num_threads <n> -out Rph_nohits123.blastn
```

6) Low-quality amino acid sequence annotation

Finally, if no annotation is obtained from the previous steps, we use HMMER and PSI-BLAST (Altschul et al. 1997) to further search for remote homology. This module is based on the concept of protein domain conservation, and includes: 1) a HMMER search against the Pfam database with more permissive thresholds than those used in other steps of the pipeline; 2) a PSI-BLAST performed nr database.

In this step we use the same input file used for the BLASTN search in step 5 (i.e.: contigs that did not get hits from the BLASTX search in step 1; contigs for which no ORFs were predicted in step 2; contigs with no amino acid sequence annotation in step 3) translated with the EMBOSS 'transeq' tool. Clearly, this last step yields a low-quality annotation, therefore a particular caution is needed in inferring homology and biological function from the results of this analysis; the annotation obtained with this module must be supported by human supervision.

As a last step, when no annotation is retrieved even through this last step, we use different tools to obtain structural feature annotation (see table 5 for an example of software that may be used for this purpose).

Appendix F

Translate sequences in six frames

```
transeq -sequence Rph_nohits123.fasta -outseq  
Rph_nohits123_allframes.fasta -alternative -frame 6 -trim T
```

HMMER

```
hmmscan -E 0.001 --domE 1 --tblout Rph_nohits123_allframes.tblout --  
domtblout Rph_nohits123_allframes.domtblout -o Rph_nohits123_allframes.out  
--noali --cpu <n> <Pfam_database.hmm> Rph_nohits123_allframes.fasta
```

PSI-Blast

```
psiblast -db nr -query Rph_nohits123_allframes.fasta -evaluate 0.1 -  
num_iterations 5 -outfmt "6 qseqid sseqid qstart qend sstart send stitle  
staxids length pident evalue " -num_threads <n> -out  
Rph_nohits123_allframes.psiblast
```

Non-annotated contigs

```
grep "Locus" Rph_nohits123_allframes.tblout | awk '{print $3}' | awk  
'BEGIN {FS = "_"} ; {print $1"_"$2}' | sort -u > Rph_hmmer_6.ids  
grep "Locus" Rph_nohits123_allframes.psiblast | cut -f1 | awk 'BEGIN {FS =  
"_"} ; {print $1"_"$2}' | sort | uniq > Rph_psiblast.ids  
comm -23 Rph_nohits123.ids Rph_nohits123_allframes_psiblast.ids >  
nopsiblast.tmp  
comm -23 nopsiblast.tmp Rph_nohits123_allframes_hmmer_6.ids >  
Rph_nohits1236.ids  
rm nopsiblast.tmp  
grep "Locus" Rph_nohits123.blastn | cut -f1 | less | sort -u >  
Rph_blastn.ids  
comm -23 Rph_nohits1236.ids Rph_blastn.ids > Rph_nohits12356.ids  
file="Rph_nohits1236.ids"; name=$(cat $file); for locus in $name; do grep  
-w -A1 $locus Rph_contigs_unwrapped.fasta; done > Rph_nohits1236.fasta  
file="Rph_nohits12356.ids"; name=$(cat $file); for locus in $name; do grep  
-w -A1 $locus Rph_contigs_unwrapped.fasta; done > Rph_nohits12356.fasta
```

ANNOTATION STEP	FEATURE	NUMBER OF CONTIGS	% OF TOTAL
—	Assembled	22818	100.00
1 - Filtering	BLASTX Hits	13112	57.5
1 - Filtering	Metazoan Hits	12733	55.8
1 - Filtering	Non-Metazoan Hits	379	1.7
2 - ORF Prediction	Predicted ORFs	12586	55.2
3 - HQ Amino Acid Annotation	Annotated	12371	54.2
3 - HQ Annotation	With GO Terms	12064	52.9
4 - Orthology	With Ortholog	12494	54.8
1 + 2 + 3	No Hits	10068	44.1
5 - Nucleotide Annotation	Annotated	3997	17.5
6 - LQ Amino Acid Annotation	Annotated	1950	8.5
All	Total Annotated	17888	78.4
All	No Annotation	4930	21.6

Table 4. Annotation results of the *R. philippinarum* transcriptome. Colors recalls those used for different steps of the annotation pipeline in Figure 7

<p>List of genomes used as BLASTX databases for findorf (1)</p> <p><i>Acyrtosiphon pisum</i> (Arthropoda, Insecta, Hemiptera) <i>Amphimedon queenslandica</i> (Porifera, Demospongiae, Haplosclerida) <i>Bombyx mori</i> (Arthropoda, Insecta, Lepidoptera) <i>Brugia malayi</i> (Nematoda, Chromadorea, Spirurida) <i>Caenorhabditis elegans</i> (Nematoda, Chromadorea, Rhabditida) <i>Capitella teleta</i> (Annelida, Polychaeta, Capitellida) <i>Crassostrea gigas</i> (Mollusca, Bivalvia, Ostreoida) <i>Drosophila melanogaster</i> (Arthropoda, Insecta, Diptera) <i>Helobdella robusta</i> (Annelida, Hirudinida, Rhynchobdellida) <i>Lingula anatina</i> (Brachiopoda, Lingulata, Lingulida) <i>Lottia gigantea</i> (Mollusca, Gastropoda, Patellogastropoda) <i>Mnemiopsis leidyi</i> (Ctenophora, Cyclocoela, Lobata) <i>Nasonia vitripennis</i> (Arthropoda, Insecta, Hymenoptera) <i>Nematostella vectensis</i> (Cnidaria, Anthozoa, Actiniaria) <i>Pinctada fucata</i> (Mollusca, Bivalvia, Pterioida) <i>Strigamia maritima</i> (Arthropoda, Chilopoda, Geophilomorpha) <i>Strongylocentrotus purpuratus</i> (Echinodermata, Echinoidea, Echinoidea) <i>Tetranychus urticae</i> (Arthropoda, Aracnida, Acariformes) <i>Trichoplax adhaerens</i> (Placozoa)</p>	<p>Databases used in the InterProScan 5 search (2)</p> <p>TIGRFAM-13.0, Panther 8.1, SMART-6.2, PrositePatterns-20.97, SuperFamily-1.75, PRINTS-42.0, Gene3d-3.5.0, PIRSF-2.84, PrositeProfiles-20.97</p>
	<p>List of the taxa used for orthology search with Proteinortho and OrthoFinder (3)</p> <p>- Peptides predicted from complete genomic sequence: <i>Capitella teleta</i> (Annelida, Polychaeta, Capitellida) <i>Crassostrea gigas</i> (Mollusca, Bivalvia, Ostreoida) <i>Helobdella robusta</i> (Annelida, Hirudinida, Rhynchobdellida) <i>Lottia gigantea</i> (Mollusca, Gastropoda, Patellogastropoda) <i>Octopus bimaculoides</i> (Mollusca, Cephalopoda, Octopoda) <i>Pinctada fucata</i> (Mollusca, Bivalvia, Pterioida) - Peptides predicted by RNA-Seq: <i>Ruditapes decussatus</i> (Mollusca, Bivalvia, Veneroidea) <i>Ruditapes philippinarum</i> (Mollusca, Bivalvia, Veneroidea)</p>
	<p>OrthoMCL DB Version 5 (4)</p> <p>List of the software utilized for protein feature annotation (5)</p> <p>SignalP, TMHMM, PSORT, COILS, PredictProtein, Jpred</p>

Table 5 . Example of databases and software used in the transcription annotation of *R. philippinarum*. Colors recalls those used for different steps of the annotation pipeline in Figure 7

Chapter 2:

The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae)

The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae)

Fabrizio Ghiselli^{1,*}, Liliana Milani^{1,*}, Mariangela Iannello¹, Emanuele Procopio¹, Peter L. Chang², Sergey V. Nuzhdin² and Marco Passamonti¹

¹ Department of Biological, Geological and Environmental Sciences, University of Bologna, Italy, Bologna, Italy

² Department of Biological Sciences, Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA

* These authors contributed equally to this work.

ABSTRACT

Despite the large number of animal complete mitochondrial genomes currently available in public databases, knowledge about mitochondrial genomics in invertebrates is uneven. This paper reports, for the first time, the complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus*, also known as the European clam. *Ruditapes decussatus* is morphologically and ecologically similar to the Manila clam *Ruditapes philippinarum*, which has been recently introduced for aquaculture in the very same habitats of *Ruditapes decussatus*, and that is replacing the native species. Currently the production of the European clam is almost insignificant, nonetheless it is considered a high value product, and therefore it is an economically important species, especially in Portugal, Spain and Italy. In this work we: (i) assembled *Ruditapes decussatus* mitochondrial genome from RNA-Seq data, and validated it by Sanger sequencing; (ii) analyzed and characterized the *Ruditapes decussatus* mitochondrial genome, comparing its features with those of other venerid bivalves; (iii) assessed mitochondrial sequence polymorphism (SP) and copy number variation (CNV) of tandem repeats across 26 samples. Despite using high-throughput approaches we did not find evidence for the presence of two sex-linked mitochondrial genomes, typical of the doubly uniparental inheritance of mitochondria, a phenomenon known in ~100 bivalve species. According to our analyses, *Ruditapes decussatus* is more genetically similar to species of the Genus *Paphia* than to the congeneric *Ruditapes philippinarum*, a finding that bolsters the already-proposed need of a taxonomic revision. We also found a quite low genetic variability across the examined samples, with few SPs and little variability of the sequences flanking the control region (Largest Unassigned Regions (LURs)). Strikingly, although we found low nucleotide variability along the entire mitochondrial genome, we observed high levels of length polymorphism in the LUR due to CNV of tandem repeats, and even a LUR length heteroplasmy in two samples. It is not clear if the lack of genetic variability in the mitochondrial genome of *Ruditapes decussatus* is a cause or an effect of the ongoing replacement of *Ruditapes decussatus* with the invasive *Ruditapes philippinarum*, and more analyses, especially on nuclear sequences, are required to assess this point.

Submitted 19 May 2017
Accepted 25 July 2017
Published 22 August 2017

Corresponding author
Fabrizio Ghiselli,
fabrizio.ghiselli@unibo.it

Academic editor
Tim Collins

Additional Information and
Declarations can be found on
page 23

DOI 10.7717/peerj.3692

© Copyright
2017 Ghiselli et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Zoology

Keywords Complete mitochondrial genome, Mitochondrial length polymorphism, Mitochondrial repeats, Codon usage, Bivalve molluscs, European clam, Comparative mitochondrial genomics, mtDNA *de novo* assembly, RNA-Seq, Doubly uniparental inheritance

INTRODUCTION

Despite a large number of animal complete mitochondrial genomes (mtDNAs) being available in public databases (>55,000 in GenBank), up to now sequencing has been focused mostly on vertebrates (~50,000 in GenBank), and the current knowledge about mitochondrial genomics in invertebrates—with the notable exception of few model organisms (e.g., *Drosophila* and *Caenorhabditis elegans*)—is uneven. To better understand invertebrate mitochondrial biology—and, most importantly, mitochondrial biology and evolution in general—it is necessary to adopt a more widespread approach in gathering and analyzing data. Failing to do so would bias our knowledge toward a few taxonomic groups, with the risk of losing a big part of the molecular and functional diversity of mitochondria. Actually, despite maintaining its core features in terms of genetic content, mtDNA in Metazoa shows a wide range of variation in some other traits such as, for example, genome architecture, abundance of unassigned regions (URs)—namely regions with no assigned product (protein, RNA)—repeat content, gene duplications, introns, UTRs, and even additional coding genes (see [Breton et al., 2014](#) for a review) or genetic elements (e.g., small RNAs, see [Pozzi et al., 2017](#)). All this emerging diversity is in sharp contrast with the—at this point outdated—textbook notion about mtDNAs role being limited to the production of a few subunits of the protein complexes involved in oxidative phosphorylation.

This paper reports, for the first time, the complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Linnaeus, 1758). *Ruditapes decussatus*—also known as the European clam—is distributed all over the Mediterranean coasts, as well as on the Atlantic shores, from Lofoten Islands (Norway) to Mauritania, including the British Isles. *Ruditapes decussatus* lives in warm coastal waters, especially in lagoons, and it is morphologically and ecologically similar to the Manila clam *Ruditapes philippinarum*, which has been recently introduced for aquaculture in the very same habitats of *Ruditapes decussatus*. *Ruditapes philippinarum*, native from the Philippines, Korea, and Japan, was accidentally introduced into North America in the 1930s, and from there it was purposely introduced in France (1972), UK (1980), and Ireland (1982) for aquaculture purposes ([Gosling, 2003](#)). According to historical records, *Ruditapes decussatus* was one of the most important species for aquaculture in Europe, but overfishing, irregular yields, recruitment failure, and outbreaks of bacterial infection pushed the producers to introduce *Ruditapes philippinarum*; Italy imported large quantities of *Ruditapes philippinarum* seed from UK in 1983 and 1984. Compared to the European clam, the Manila clam turned out to be faster growing, more resistant to disease, to have a more extended breeding period and a greater number of spawning events, and to begin sexual maturation earlier (i.e., at a smaller size). Upon introduction of the more robust *Ruditapes philippinarum*, *Ruditapes decussatus* suffered a population decline in the Southwestern Europe ([Arias-Pérez et al., 2016](#)), and

currently the production of the European clam is almost insignificant. Nonetheless the grooved carpet shell is considered a high value product, and therefore it is an economically important species, especially in Portugal, Spain and Italy ([Gosling, 2003](#); [Leite et al., 2013](#); [de Sousa et al., 2014](#)).

Molluscs in general, and bivalves in particular, exhibit an extraordinary degree of mtDNA variability and unusual features, such as: large mitochondrial genomes (up to ~47Kb), high proportion of URs (i.e., number of base pairs annotated as URs over the total mtDNA length), novel protein coding genes with unknown function, frequent and extensive gene rearrangement, and differences in strand usage ([Gissi, Iannelli & Pesole, 2008](#); [Breton et al., 2011](#); [Ghiselli et al., 2013](#); [Milani et al., 2014b](#); [Plazzi, Puccio & Passamonti, 2016](#)). Moreover, mitochondrial genome size varies among bivalves because of gene duplications and losses ([Serb & Lydeard, 2003](#); [Passamonti et al., 2011](#); [Ghiselli et al., 2013](#)), and sometimes genes are fragmented as in the case of ribosomal genes in oysters ([Milbury et al., 2010](#)). The most notable feature of bivalve mtDNA is the doubly uniparental inheritance (DUI) system of transmission ([Skibinski, Gallagher & Beynon, 1994a, 1994b](#); [Zouros et al., 1994a, 1994b](#)). Under DUI, two different mitochondrial lineages (and their respective genomes) are transmitted to the progeny: one is inherited from the egg (female-transmitted or F-type mtDNA), the other is inherited from the spermatozoon (male-transmitted or M-type mtDNA). Following fertilization, the early embryo is heteroplasmic, but the type of mitochondria present in the adult is tightly linked to its sex. Females are commonly homoplasmic for F, while males are heteroplasmic with the following distribution of mtDNA types: the germ line is homoplasmic for the M-type (which will be transmitted via sperm to male progeny), the soma is heteroplasmic to various degrees, depending on tissue type and/or species ([Ghiselli, Milani & Passamonti, 2011](#); [Zouros, 2013](#)). To date, the only known animals exhibiting DUI are about 100 species of bivalve molluscs ([Gusman et al., 2016](#)). This natural and evolutionarily stable heteroplasmic system can be extremely useful to investigate several aspects of mitochondrial biology (see [Passamonti & Ghiselli, 2009](#); [Breton et al., 2014](#); [Milani & Ghiselli, 2015](#); [Milani, Ghiselli & Passamonti, 2016](#)). Indeed, despite the fact that many aspects of DUI are still unknown, there is evidence that DUI evolved from a strictly maternal inheritance (SMI) system ([Milani & Ghiselli, 2015](#); [Milani, Ghiselli & Passamonti, 2016](#)), by modifications of the molecular machinery involved in mitochondrial inheritance, through as-yet-unknown specific factors (see [Diz, Dudley & Skibinski, 2012](#); [Zouros, 2013](#) for proposed models). The detection of DUI is not a straightforward process, especially using PCR-based approaches: given that the divergence between F and M genomes is often comparable to the distance between mtDNAs of different classes of Vertebrates, primers may fail to amplify one of the two mtDNAs, yielding a false-negative result. Moreover, M-type mtDNA can be rare in somatic tissues, so it may be difficult to amplify from animals sampled outside of the reproductive season, when gonads are absent (thoroughly discussed in [Theologidis et al., 2008](#)). High-throughput sequencing (HTS) approaches can overcome such problems, because a prior knowledge of the mtDNA sequence is not needed, and low-copy variants can be easily unveiled (see [Ju et al., 2011](#); [King et al., 2014](#)). Until now, HTS has been scarcely

utilized to study mitochondrial transcriptomes and genomes ([Pesole et al., 2012](#); [Smith, 2013](#)), even if it showed very good potential ([Lubošny et al., 2017/2](#); [Yuan et al., 2016](#)). In this work we: (i) assembled *Ruditapes decussatus* mitochondrial genome from RNA-Seq data, and validated it by Sanger sequencing, (ii) analyzed and characterized *Ruditapes decussatus* mitochondrial genome, comparing its features with those of other venerid bivalves; (iii) assessed mitochondrial sequence polymorphism (SP) and structural variants—copy number variation (CNV) of tandem repeats—among the sampled animals.

MATERIALS AND METHODS

Sampling

The 26 *Ruditapes decussatus* specimens used in this study were collected from the Northern Adriatic Sea, in the river Po Delta region (Sacca di Goro, approximate GPS coordinates: 44°50'06"N, 12°17'55"E) at the end of July 2011, during the spawning season. Each individual was dissected, and gonadal liquid collected with a glass capillary tube. All the samples showed ripe gonads, consistently with the time of the year when the sampling occurred. The gonadal liquid was checked under a light microscope to assess the sex of the individual, and to make sure that the sample consisted of mature gametes. Both the gamete samples and the clam bodies were flash-frozen in liquid nitrogen, and stored at -80°C , until nucleic acid extraction. [Table S1](#) shows the sample list, and details about data availability.

RNA-Seq

In total, 12 samples (six males and six females) were used for RNA-Seq. Total RNA extraction and library preparation were performed following the protocol described in [Mortazavi et al. \(2008\)](#), with the modifications specified in [Ghiselli et al. \(2012\)](#). The 12 samples were indexed, pooled and sequenced in two lanes (two technical replicates) of Illumina GA IIX, using 76 bp paired-end reads.

De novo assembly

The mitochondrial genome of *Ruditapes decussatus* was not available in the databases, so we used the transcriptome data to generate a draft to be used as a guide for Sanger sequencing. Illumina reads from all 12 samples were pooled and compared to a set of 20 Bivalvia mitochondrial genomes to identify reads with mitochondrial origin. Alignment was done using BLASTN. All reads with similarity yielding $E\text{-value} < 1\text{E-}5$ were then assembled into contigs using the A5 pipeline (version 2013032; [Tritt et al., 2012](#)) and joined into scaffolds using CAP3 ([Huang & Madan, 1999](#)). For the quality check step, we applied a PHRED Q-score cutoff threshold of 33; the other A5 parameters were set as default. CAP3 was run with default settings as well.

Sanger validation

In total, 14 *Ruditapes decussatus* samples from the same collection campaign—sexed, and stored at -80°C —were used for DNA extraction. DNA from the gonadic tissue was

extracted using the Qiagen DNeasy kit. Primers for mtDNA amplification were designed based on contigs obtained from RNA-Seq matching venerid mtDNA sequences, then the “primer walking” method was used to Sanger-sequence the complete mitochondrial genome of *Ruditapes decussatus*. The primers were designed with the software Primer3 (Rozen & Skaletsky, 2000) and tested on several samples, then a female was chosen as reference sample for Sanger validation of mtDNA *de novo* assembly. In addition, we amplified the largest unassigned region (LUR) of 13 females to assess its variability (see Results and Discussion). The list of the primers and their sequences are reported in Table S2. PCR reactions were performed in a final volume of 50 μ l using the GoTaq Flexi DNA Polymerase Kit (Promega, Madison, WI, USA), on a 2720 Thermal Cycler (Applied Biosystem, Foster City, CA, USA). The PCR reactions were set as follows: initial denaturation 95 °C for 1 min, then 30 cycles of amplification (denaturation 95 °C for 1 min, annealing 48–60 °C for 1 min, extension 72 °C for 1 min/kb), then the final extension at 72 °C for 5 min. PCR products were checked by electrophoretic run on 1% agarose gel, and then purified using the DNA Clean & Concentrator-25 kit (Zymo Research, Irvine, CA, USA).

Sanger sequencing was performed by Macrogen Inc. (<http://www.macrogen.com>).

Sequences were aligned with the software MEGA 6.0 (Tamura et al., 2013), using the contigs obtained by RNA-seq as a reference.

Annotation

Open reading frames (ORFs) were identified with ORF finder (Wheeler et al., 2005). Alternative start codons were considered functional because they are common in Bivalvia. ORFs were annotated starting from the first available start codon (ATG, ATA, or ATC) downstream of the preceding gene, and ending with the first stop codon in frame (TAA or TAG). tRNA genes and their structure were identified with MITOS (Bernt et al., 2013) and ARWEN (Laslett & Canback, 2008). Secondary structures were predicted using the RNAFold Server, included in the ViennaRNA Web Services (<http://rna.tbi.univie.ac.at/>; Gruber et al., 2008); the folding temperature was set at 16 °C which is the average annual temperature of the water from which the *Ruditapes decussatus* specimens used in this work were fished (download RNAFold results from figshare: <https://ndownloader.figshare.com/files/8387672>). tRNAs and other secondary structures were drawn with the software Varna GUI (Darty, Denise & Ponty, 2009). Ribosomal small subunit (*rrnS*) and large subunit (*rrnL*) were identified with BLASTN, and annotated considering the start and the end of the adjacent genes as the boundaries of the rRNA genes. Non-genic regions were annotated as URs. In order to identify the putative D-loop/control region (CR), we analyzed the LUR with the MEME suite (Bailey et al., 2009) to find DNA motifs using the following bivalve species as comparison: *Acanthocardia tuberculata*, *Arctica islandica*, *Coelomactra antiquata*, *Fulvia mutica*, *Hiatella arctica*, *Loripes lacteus*, *Lucinella divaricata*, *Lutraria rhynchaena*, *Meretrix lamarckii* (F-type), *Meretrix lamarckii* (M-type), *Meretrix lusoria*, *Meretrix lyrata*, *Meretrix meretrix*, *Meretrix petechialis*, *Moerella iridescens*, *Nuttallia olivacea*, *Paphia amabilis*, *Paphia euglypta*, *Paphia textile*, *Paphia undulata*, *Ruditapes philippinarum* (F-type), *Ruditapes philippinarum* (M-type),

Semele scabra, *Sinonovacula constricta*, *Solecurtus divaricatus*, *Solen grandis*, *Solen strictus*, *Soletellina diphos* and the sea urchin *Strongylocentrotus purpuratus* (Echinoidea, Strongylocentrotidae). The list of the species used in the phylogenetic analysis and in the comparative analyses of DNA motifs, sequence similarity, and gene order are available in [Table S3](#). The GOMo (Gene Ontology for Motifs; [Buske et al., 2010](#)) tool of the MEME suite was used to assign GO terms to the motifs discovered.

The number of repeats in the LUR of the reference sample (F4) was calculated with tandem repeat finder (<http://tandem.bu.edu/trf/trf.html>), since the complete LUR sequence was available (download tandem repeat finder results from Figshare: <https://ndownloader.figshare.com/files/8387666>). In the other cases, in which the LUR could not be sequenced without gaps, the number of repeats was inferred from agarose gel electrophoresis.

Other analyses

Comparisons among venerid complete mtDNAs were performed with BLAST Ring Image Generator (BRIG) ([Alikhan et al., 2011](#)) and Easyfig ([Sullivan, Petty & Beatson, 2011](#)). Descriptive statistics were obtained with MEGA v6.0 ([Tamura et al., 2013](#)), except for the codon usage table, which was obtained with the Sequence Manipulation Suite ([Stothard, 2000](#)). SP assessment from RNA-Seq reads was performed with the Genome Analysis Toolkit (GATK, [McKenna et al., 2010](#)), with the Sanger-sequenced mtDNA as reference. For SP discovery and genotyping we used standard hard filtering parameters or variant quality score recalibration ([DePristo et al., 2011](#)). The MitoPhast pipeline ([Tan et al., 2015](#)) was used to obtain the Maximum Likelihood (ML) tree, which was visualized with Evolview v2 ([He et al., 2016](#)). Briefly, MitoPhast takes as input GenBank files (.gb), extracts the coding sequences, profiles the sequences with Pfam ([Finn et al., 2016](#)) and PRINTS ([Attwood et al., 2003](#)), performs a multiple sequence alignment with Clustal Omega ([Sievers et al., 2011](#)), removes poorly aligned regions with trimAl ([Capella-Gutiérrez, Silla-Martinez & Gabaldón, 2009](#)), concatenates the coding sequences, performs data partitioning and model selection, and then carries out a ML analysis using RAxML ([Stamatakis, 2014](#)). The species used in the ML analysis, and their GenBank Accession Numbers are listed in [Table S3](#). Amino acid sequences of three different *cox3* ORFs inferred from Sanger sequencing and GATK polymorphism data were analyzed with InterProScan ([Jones et al., 2014](#)).

RESULTS

De novo assembly and Sanger validation

Despite using HTS on extracts of ripe gonads (i.e., mature gametes), and multiple assembly strategies (see Discussion for details) we could not find evidence for DUI. The de novo assembly process produced 9 contigs, of which 8 included multiple genes, and one included a single gene (see [Table 1](#)). The sequences of the contigs in FASTA format are available on figshare (<https://ndownloader.figshare.com/files/8906839>). In four cases (Contigs 1, 3, 6, and 7) a clear polyadenylation signal was present, in other four cases (Contigs 2, 5, 8, and 9) it was not. Contig 4, the only one including a single gene (*cox3*), ends with just 8 As, so it is not clear if a polyadenylation signal is present in this case.

Table 1 Features of the contigs obtained by *de novo* assembly of mtDNA.

Contig	Length	Gene content	Poly-A	Notes
1	6,794	<i>atp6_nd3_nd5_cox1_</i> <i>tRNA-Leu1_nd1_nd2_nd4L</i>	Yes	Chimeric assembly. The contiguity between <i>nd5</i> and <i>cox1</i> is an artifact
2	1,884	<i>rrnS_cox3</i>	No	–
3	1,288	<i>atp6_nd3</i>	Yes	–
4	1,663	<i>cox3</i>	?	The contig ends with just 8 As
5	1,934	<i>atp8_nd4_tRNA-His_tRNA-Glu_</i> <i>tRNA-Ser2_tRNA-Tyr</i>	No	–
6	1,831	<i>atp8_nd4_tRNA-His</i>	Yes	–
7	5,478	<i>cox2_tRNA-Ile_nd4L_nd2_nd1_</i> <i>tRNA-Leu1_cox1</i>	Yes	There is a polyadenylation signal (56 As) after the <i>cox2</i> gene
8	2,879	<i>cytb_rrnL</i>	No	–
9	952	<i>nd6_tRNA-Lys_tRNA-Val_tRNA-Phe_</i> <i>tRNA-Trp_tRNA-Arg_tRNA-Leu2</i>	No	–

In Contig 7 (that includes *cox2*, *tRNA-Ile*, *nd4L*, *nd2*, *nd1*, *tRNA-Leu1*, and *cox1* genes) there is a polyadenylation signal (56 As) after the *cox2* gene.

The nine contigs were used as a scaffold for the primer walking procedure used for Sanger validation of the *de novo* assembly. We first tried to connect the contigs designing primers close to the 5' and 3' ends of each contig and pairing them following the gene order of *Paphia*, because the sequence of genes in the contigs suggested that *Ruditapes decussatus* gene order might have been similar. During such process, Contig 1 turned out to be a chimeric assembly between two non-contiguous portions of the mtDNA, one including *atp6*, *nd3*, and *nd5*, the other including *cox1*, *tRNA-Leu1*, *nd1*, *nd2*, and *nd4L*. Once we amplified and sequenced the portions of mtDNA between the contigs, we proceeded with the Sanger resequencing of the remaining parts.

Annotation and mtDNA features

The mitochondrial genome contains 13 protein-coding genes, and in the reference female is 18,995 bp long (Fig. 1); the gene arrangement and other details are shown in Table 2. All genes are located on the heavy strand, and in addition to the classic start codon ATG (Met), the alternative start codons ATA (Met) and ATC (Ile) are present. The most frequently used start codons are: ATA (*cox1*, *nd1*, *nd4L*, *cox2*, *cob*, *atp8*, *nd4*), and ATG (*nd2*, *atp6*, *nd3*, *nd5*, *nd6*, *cox3*). The stop codons found are TAG (*cox1*, *nd2*, *nd4L*, *cox2*, *cytb*, *nd4*) and TAA (*nd1*, *atp6*, *nd3*, *atp8*, *nd6*). The *nd4* gene has an incomplete stop codon (TA-). 22 tRNA genes were identified, including two tRNAs for leucine, tRNA-Leu1 (TAG) and tRNA-Leu2(TAA), and two for serine, tRNA-Ser1(TCT) and tRNA-Ser2 (TGA), both showing degenerate D-arm branches. tRNA structures are shown in Fig. S1. The two rRNAs, *rrnS* and *rrnL*, were both identified: the *rrnS* is located between *cox3* and *cox1*, while *rrnL* is between *cytb* and *atp6*. URs were identified on the basis of unannotated spaces between different genes; we found 24 URs (Table 3).

The analysis of the nucleotide composition points out that the mitochondrial genome of this bivalve species exhibits high A+T content, totaling 63% vs 37% G+C. The

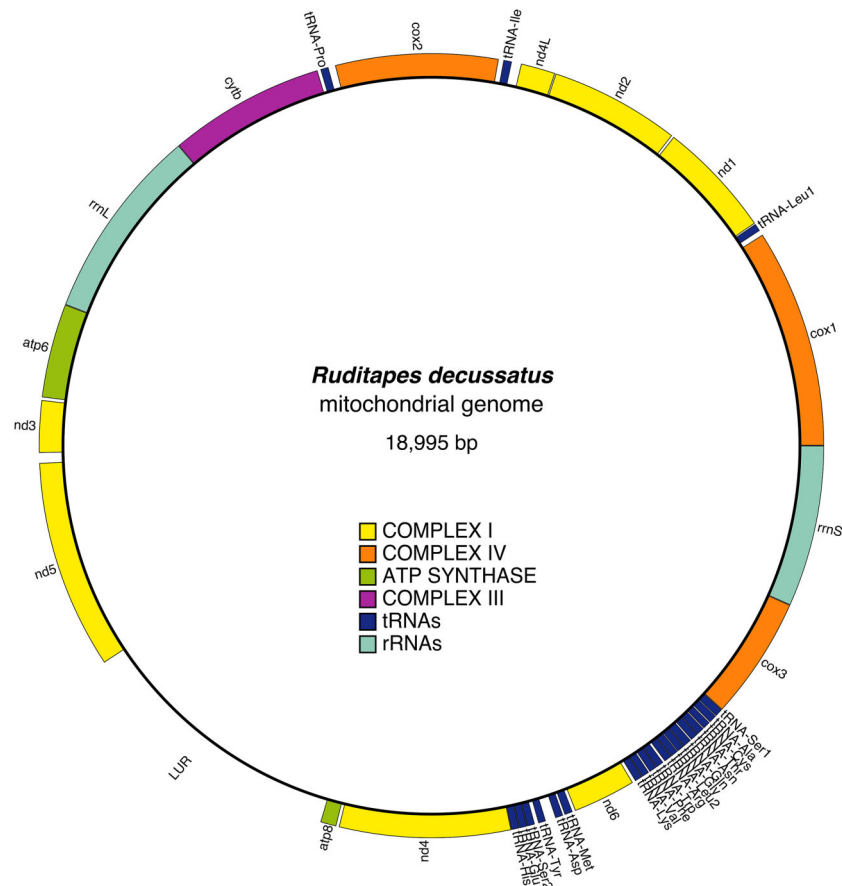


Figure 1 *Ruditapes decussatus* mtDNA gene arrangement.

minimum values of A+T are found in *cytb* (60.1%) and *nd4* (61%). The nucleotide composition of every gene is shown in Table 4. According to the analysis above, both A and T occur very frequently at the third position of codons (64.6% on average of A+T), while the less frequent base in third position is C (12%). The most used codons are UUU (Phe), counted 269 times, and UUA (Leu) counted 210 times (6.78% and 5.29% of the total, respectively), while the less used codons are CGC (Arg) counted 6 times (0.15%), ACC (Thr) and CCG (Pro) each counted 16 times (0.4%) (Table 5). Only in four cases over 20 (Lys, Leu, Gln, Val), the most frequently used codon matches the correspondent mitochondrial tRNA anticodon.

The UR11 is the LUR and is located between *atp8* and *nd5* (Figs. 1 and 2A). The LUR of the female used for whole mtDNA Sanger sequencing (i.e., the reference female, F4) is 2,110 bp long, and includes 6.5 repeated sequences—each repeat having a length of 54 bp—localized in the 3' region of the LUR, just upstream the *atp8* gene (Fig. 2A). DNA secondary structure analysis predicted three stem-loop structures in such region (Fig. 2B and Supplemental Information files on figshare: <https://ndownloader.figshare.com/files/8387672>), with a change in Gibbs free energy (ΔG) of -71.38 Kcal/mol. We also amplified and sequenced the LUR of 13 more females. We were not able to completely sequence LURs longer than 2,110 bp, because of the known difficulties in Sanger sequencing of

Table 2 MtDNA gene arrangement of *Ruditapes decussatus*.

Name	Type	Start	Stop	Length (bp)	Start	Stop	Anticodon
<i>cox1</i>	Coding	1	1,716	1,716	ATA	TAG	
<i>tRNA-Leu1</i>	tRNA	1,754	1,815	62			TAG
<i>nd1</i>	Coding	1,822	2,739	918	ATA	TAA	
<i>nd2</i>	Coding	2,755	3,774	1,020	ATG	TAG	
<i>nd4l</i>	Coding	3,780	4,052	273	ATA	TAG	
<i>tRNA-Ile</i>	tRNA	4,125	4,190	66			GAT
<i>cox2</i>	Coding	4,228	5,499	1,272	ATA	TAG	
<i>tRNA-Pro</i>	tRNA	5,553	5,616	64			TGG
<i>cytb</i>	Coding	5,641	6,864	1,224	ATA	TAG	
<i>rrnL</i>	rRNA	6,865	8,385	1,521			
<i>atp6</i>	Coding	8,386	9,123	738	ATG	TAA	
<i>nd3</i>	Coding	9,145	9,552	408	ATG	TAA	
<i>nd5</i>	Coding	9,631	11,268	1,638	ATG	TAG	
<i>atp8</i>	Coding	13,379	13,504	126	ATA	TAA	
<i>nd4</i>	Coding	13,526	14,865	1,340	ATA	TA-	
<i>tRNA-His</i>	tRNA	14,866	14,928	63			GTG
<i>tRNA-Glu</i>	tRNA	14,929	14,990	62			TTC
<i>tRNA-Ser2</i>	tRNA	14,991	15,052	62			TGA
<i>tRNA-Tyr</i>	tRNA	15,081	15,140	60			GTA
<i>tRNA-Asp</i>	tRNA	15,218	15,280	63			GTC
<i>tRNA-Met</i>	tRNA	15,294	15,358	65			CAT
<i>nd6</i>	Coding	15,380	15,874	495	ATG	TAA	
<i>tRNA-Lys</i>	tRNA	15,897	15,959	63			TTT
<i>tRNA-Val</i>	tRNA	15,960	16,021	62			TAC
<i>tRNA-Phe</i>	tRNA	16,030	16,092	63			GAA
<i>tRNA-Trp</i>	tRNA	16,093	16,155	63			TCA
<i>tRNA-Arg</i>	tRNA	16,171	16,232	62			TCG
<i>tRNA-Leu2</i>	tRNA	16,233	16,295	63			TAA
<i>tRNA-Gly</i>	tRNA	16,297	16,358	62			TCC
<i>tRNA-Gln</i>	tRNA	16,359	16,427	69			TTG
<i>tRNA-Asn</i>	tRNA	16,435	16,497	63			GTT
<i>tRNA-Thr</i>	tRNA	16,498	16,560	63			TGT
<i>tRNA-Cys</i>	tRNA	16,565	16,626	62			GCA
<i>tRNA-Ala</i>	tRNA	16,632	16,696	65			TGC
<i>tRNA-Ser1</i>	tRNA	16,698	16,764	67			TCT
<i>cox3</i>	Coding	16,765	17,730	966	ATG	TAA	
<i>rrnS</i>	rRNA	17,731	18,995	1,265			

Note:

The anticodon of tRNAs are reported in the 5'-3' direction.

Table 3 Unassigned regions (URs).

UR name	Start	Stop	Length (bp)
UR1	1,717	1,753	37
UR2	1,816	1,821	6
UR3	2,740	2,754	15
UR4	3,775	3,779	5
UR5	4,053	4,124	72
UR6	4,191	4,227	37
UR7	5,500	5,552	53
UR8	5,617	5,640	24
UR9	9,124	9,144	21
UR10	9,553	9,630	78
UR11 (LUR)	11,269	13,378	2,110
UR12	13,505	13,525	21
UR13	15,053	15,080	28
UR14	15,141	15,217	77
UR15	15,281	15,293	13
UR16	15,359	15,379	21
UR17	15,875	15,896	22
UR18	16,022	16,029	8
UR19	16,156	16,170	15
UR20	16,296	16,296	1
UR21	16,428	16,434	7
UR22	16,561	16,564	4
UR23	16,627	16,631	5
UR24	16,697	16,697	1

regions including multiple repeats. The sequence alignment of the 13 LURs is available for download from figshare (<https://ndownloader.figshare.com/files/8360789>). LUR lengths, inferred from gel electrophoresis, are reported in Table 6, and they range from 2,000 to 5,000 bp. Two females (F3 and F17) showed length heteroplasmy of the LUR. The portion of the genome occupied by URs varies between 14.11% and 29.38%, depending on LUR length. The analysis with MEME (output shown in Figs. S2 and S3) unveiled two motifs (Fig. 2C) that show a strong conservation within the Veneridae family, and with *S. purpuratus*. The sea urchin was included in the analysis because Cao et al. (2004) reported a match between some motifs found in the CR of the marine mussels *Mytilus edulis* and *Mytilus galloprovincialis* with regulatory elements of the sea urchin CR. Accordingly, the search with GOMo assigned a series of GO terms related to transcription to the two motifs (Table S4).

Polymorphism

Table 7 (top) shows the statistics associated with the SP analysis performed with GATK on the 12 samples used for RNA-Seq, with the Sanger-sequenced mtDNA as reference. Overall, 257 SPs were called, of which 145 (56.4%) were located in coding sequences

Table 4 Nucleotide composition.

Name	Length (bp)	T (%)	C (%)	A (%)	G (%)	A+T (%)	T3 (%)	C3 (%)	A3 (%)	G3 (%)	A3+T3 (%)
<i>cox1</i>	1,716	35.8	15.5	25.8	22.9	61.6	39	12.1	28.0	21.3	67.0
<i>nd1</i>	918	38.7	12.5	24.0	24.8	62.7	38	10.1	30.7	21.2	68.7
<i>nd2</i>	1,020	38.3	11.0	24.8	25.9	63.1	35	11.5	29.4	24.4	64.4
<i>nd4l</i>	273	39.9	12.8	25.3	22.0	65.2	34	14.3	30.8	20.9	64.8
<i>cox2</i>	1,272	29.7	14.8	29.1	26.4	58.8	30	15.3	27.4	27.6	57.4
<i>cob</i>	1,224	37.4	17.2	22.7	22.6	60.1	41	14.7	21.8	22.1	62.8
<i>rrnL</i>	1,749	33.2	11.5	32.6	22.6	65.8	33	10.6	33.4	23.0	66.4
<i>atp6</i>	510	42.0	15.7	20.8	21.6	62.8	45	13.5	21.8	20.0	66.8
<i>nd3</i>	408	39.5	11.0	24.8	24.8	64.3	33	11.0	30.1	25.7	63.1
<i>nd5</i>	1,638	37.6	11.7	27.7	23.0	65.3	35	11.0	34.2	19.8	69.2
<i>atp8</i>	126	44.4	11.9	19.0	24.6	63.4	45	4.8	23.8	26.2	68.8
<i>nd4</i>	1,340	38.9	12.9	22.1	26.1	61.0	41	10.8	24.9	23.5	65.9
<i>nd6</i>	495	39.2	12.1	23.0	25.7	62.2	38	13.9	27.9	20.0	65.9
<i>cox3</i>	966	36.9	12.7	24.8	25.6	61.7	39	9.6	28.6	23.0	67.6
<i>rrnS</i>	1,265	32.7	12.3	32.9	22.1	65.6	35	13.5	31.6	19.5	66.6
All coding	14,920	36.3	13.2	26.5	24.0	63.0	37	12.0	28.9	22.4	65.7
All <i>rRNAs</i>	3,014	32.9	23.8	32.7	22.3	65.7					
All <i>tRNAs</i>	1,394	35.4	12.8	30.2	21.7	65.6					
All URs	2,681	28.2	14.1	34.1	23.6	62.3					
All genic DNA	16,314	36.2	13.2	26.8	23.8	63.0					
All DNA	18,995	35.1	13.3	27.9	23.7	63.0					

Note:

URs, unassigned regions.

(CDS). Interestingly, most of the SPs were called because of private alleles of one single male specimen (mRDI01). More in detail, 151 SPs out of 257 (58.7%) along the whole mtDNA sequence, and 103 SPs out of 145 (71%) in CDS, were private of mRDI01. In CDS, if we exclude the SPs associated with this male, the number of polymorphisms drops to 42 over 14,920 bp of coding mtDNA (GATK output in VCF format and a detailed list of SPs in tabular format is available on figshare: <https://ndownloader.figshare.com/files/8902537>), of which 18 are represented by indels, 6 of which are located in 4 different coding genes: one each in *cox1*, *cytb*, and *nd5*, plus 3 in *cox3* (see Table 8). A file showing the ORF generated by the different variants of *cox3*, and alignments between them is available on figshare (<https://ndownloader.figshare.com/files/8402471>). Table 7 (bottom) shows the number of SPs in males, in males except mRDI01, and in females both along the whole mtDNA, and in CDS. The number in brackets represent the number of private SPs for each category.

Comparison with other veneridae

Figure 3 shows the *Ruditapes decussatus* mtDNA map (external gray circle), and the BLASTN identity (colored inner circles) with complete mtDNAs of other 10 venerid species (see list in Table S3). Figure 4 shows the ML tree obtained with the MitoPhast

Table 5 Codon usage.

Amino acid	Codon	#	Frequency	%TOT	Amino acid	Codon	#	Frequency	%TOT
Ala	GCG	29	0.15	0.73	Pro	CCG	16	0.12	0.40
	<u>GCA</u>	44	0.23	1.11		<u>CCA</u>	36	0.27	0.91
	GCT	85	0.45	2.14		CCT	58	0.43	1.46
	GCC	30	0.16	0.76		CCC	24	0.18	0.61
Cys	TGT	94	0.76	2.37	Gln	CAG	25	0.44	0.63
	<u>TGC</u>	30	0.24	0.76		<u>CAA</u>	32	0.56	0.81
Asp	GAT	54	0.66	1.36	Arg	CGG	23	0.31	0.58
	<u>GAC</u>	28	0.34	0.71		<u>CGA</u>	21	0.28	0.53
Glu	GAG	87	0.6	2.19		CGT	25	0.33	0.63
	<u>GAA</u>	58	0.4	1.46		CGC	6	0.08	0.15
Phe	TTT	269	0.78	6.78	Ser	AGG	69	0.19	1.74
	<u>TTC</u>	78	0.22	1.97		<u>AGA</u>	69	0.19	1.74
Gly	GGG	131	0.4	3.30		AGT	55	0.15	1.39
	<u>GGA</u>	61	0.19	1.54		AGC	23	0.06	0.58
	GGT	98	0.3	2.47		TCG	18	0.05	0.45
	GGC	36	0.11	0.91		<u>TCA</u>	33	0.09	0.83
His	CAT	37	0.62	0.93		TCT	76	0.21	1.92
	<u>CAC</u>	23	0.38	0.58		TCC	22	0.06	0.55
Ile	ATT	165	0.8	4.16	Thr	ACG	21	0.17	0.53
	<u>ATC</u>	40	0.2	1.01		<u>ACA</u>	30	0.24	0.76
Lys	AAG	61	0.41	1.54		ACT	57	0.46	1.44
	<u>AAA</u>	87	0.59	2.19		ACC	16	0.13	0.40
Leu	TTG	122	0.23	3.08	Val	GTG	113	0.3	2.85
	<u>TTA</u>	210	0.39	5.29		<u>GTA</u>	121	0.32	3.05
	CTG	43	0.08	1.08		GTT	119	0.32	3.00
	<u>CTA</u>	70	0.13	1.76		GTC	23	0.06	0.58
	CTT	75	0.14	1.89		Trp	TGG	58	0.54
CTC	20	0.04	0.50	<u>TGA</u>	49		0.46	1.24	
Met	<u>ATG</u>	86	0.36	2.17	Tyr	TAT	103	0.69	2.60
	ATA	155	0.64	3.91		<u>TAC</u>	47	0.31	1.18
Asn	AAT	76	0.66	1.92	STOP	TAG	34	0.58	0.86
	<u>AAC</u>	39	0.34	0.98		TAA	25	0.42	0.63

Note:

The codons corresponding to a tRNA present in the mitochondrial genome are underlined and in bold. The highest frequency among synonymous codons is also underlined and in bold. #, number of codons; Frequency, frequency of each codon among synonymous codons; %TOT, frequency of each codon among all the codons.

pipeline; the complete input and output of this analysis is available on figshare (<https://ndownloader.figshare.com/files/8360792>). Figure 5 shows the variation in gene order between *Ruditapes decussatus* and *P. euglypta* (Fig. 5A), *M. lamarckii* F-type (Fig. 5B), *Ruditapes philippinarum* F-type (Fig. 5C), and among all the four species (Fig. 5D).

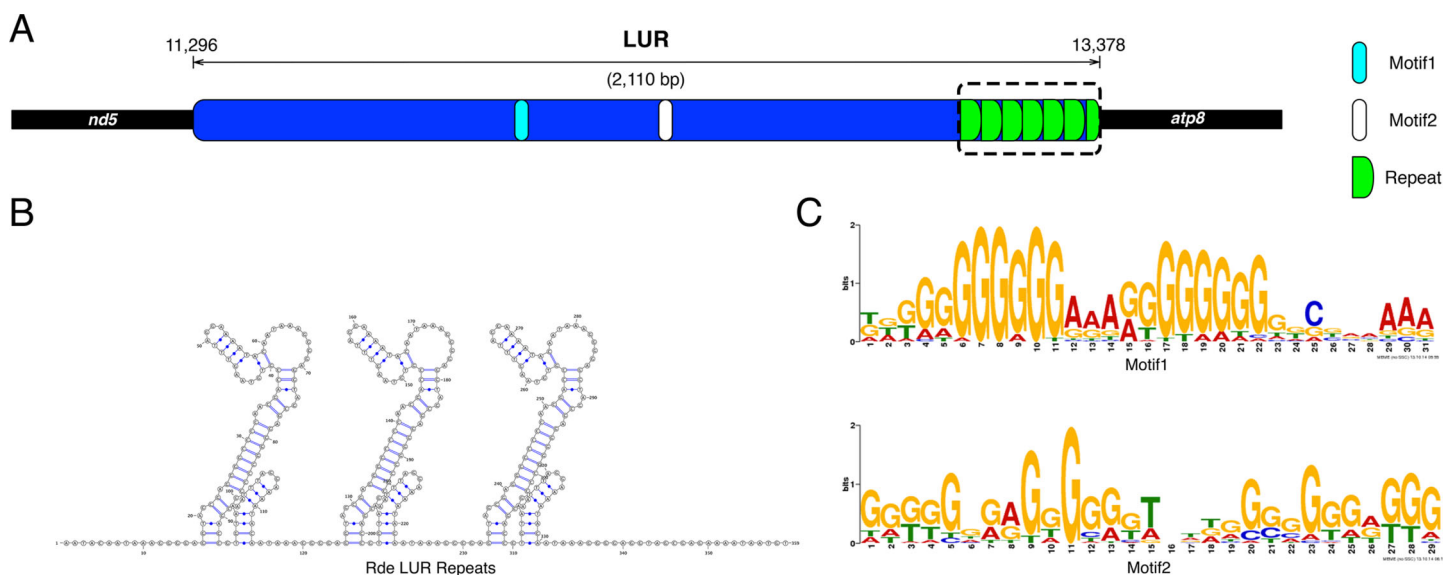


Figure 2 Principal features of the Largest Unassigned Region (LUR). (A): map of the LUR; (B): DNA secondary structure predicted in the repeat region (boxed in A); (C): Logos of the two DNA motifs found in the LUR.

Table 6 LUR length and number of repeats in the 13 female samples analyzed.

Specimen	Length (bp)	Number of repeats	GenBank Acc. No.
F3	2,100–3,500	6.5–25	MF055702
F5	5,000	45	MF055703
F7	3,500	25	MF055704
F9	3,500	25	MF055705
F10	3,000	20	MF055706
F11	3,000	20	MF055707
F13	3,500	25	MF055708
F15	3,000	20	MF055709
F16	3,500	25	MF055710
F17	2,500–3,500	8–25	MF055711
F19	3,500	25	MF055712
F20	2,500	8	MF055713
F21	2,100	6.5	MF055714

Note:

F3 and F17 are heteroplasmic with LURs of different length.

DISCUSSION

RNA-Seq-guided sequencing of mtDNA

The *de novo* assembly of the mtDNA from RNA-Seq data turned out to be informative, simplifying the primer walking procedure used for Sanger sequencing. Only one contig (Contig 1) resulted to be a chimeric sequence obtained by the misassembly of two smaller contigs. Most of the contigs (eight out of nine) contained more than one gene, and most of the tRNA genes were included in the *de novo* assembly. Except for *tRNA-Pro*,

Table 7 Sequence Polymorphism (SP): SPs and small indels called by GATK.

Feature	Value	Min	Median	Mean	Max
Depth (all SPs)	–	6	1,357	1,521	3,880
Phred score (all SPs)	–	3.30E+01	5.76E+03	4.18E+07	2.15E+09
Depth (SPs in CDS)	–	222	2,038	2,150	3,880
Phred score (SPs in CDS)	–	1.18E+02	1.01E+04	4.45E+07	2.15E+09
Total number of SPs	257	–	–	–	–
Number of mRDI01 private SPs	151 (58.7% of the total)	–	–	–	–
Number of SPs in CDS	145 (56.4% of the total)	–	–	–	–
Number of mRDI01 private SPs in CDS	103 (71% of the SP in CDS)	–	–	–	–
Number of SPs in CDS (excluding mRDI01)	42	–	–	–	–
Frequency of SPs in CDS	0.0097 (~1 every 103 bp)	–	–	–	–
Frequency of SPs in CDS (excluding mRDI01)	0.0028 (~1 every 355 bp)	–	–	–	–
Total number of indels	18	–	–	–	–
Number of indels in CDS	6	–	–	–	–
Number of indels causing frameshift	4	–	–	–	–
# Of SPs	Whole mtDNA	CDS			
Males	234 (160)	136 (107)			
Males (no mRDI01)	84 (15)	32 (6)			
Females	97 (23)	38 (9)			

Note:

CDS, coding sequences; Whole mtDNA, polymorphism in the whole mitochondrial genome; the number in brackets the bottom of the table represent private SPs (e.g., there are 23 female specific SPs in the whole mtDNA and 9 female specific SPs in CDS); *p*-value, significance of the Fisher's exact test on number of SPs between sexes (i.e., all males vs females, males except mRDI01 vs females).

Table 8 Indels located in coding sequences.

Position	Depth	Qual	Gene	SP	Frameshift	Sample	Allele frequency	Notes
1,698	3,732	1.38E+04	<i>cox1</i>	C/CAAA	No	mRDI02, mRDI03	0.089, 0.85	Insertion of 1 Lysine
6,364	1,929	2.15E+09	<i>cytb</i>	CT/C	Yes	fRDI04, mRDI05	0.80, 0.81	Yields a shorter <i>Cytb</i> . Possible sequencing error due to the homopolymer CTTTTTTT
10,449	1,780	2.15E+09	<i>nd5</i>	C/CT	Yes	fRDI01, fRDI04, fRDI05	0.11, 0.10, 0.11	Yields a <i>nd5</i> gene divided in 2 ORFs. Possible sequencing error due to the homopolymer CTTTTTTT
17,619	2,272	5.98E+03	<i>cox3</i>	AGCG/A	No	mRDI01	0.97	Deletion of one Alanine
17,621	2,188	9.99E+04	<i>cox3</i>	CG/C	Yes	mRDI01	0.99	Always combined with SP_17624. Together change the last 35 amino acids
17,624	2,287	5.98E+03	<i>cox3</i>	C/CAT	Yes	mRDI01	0.99	Always combined with SP_17621. Together change the last 35 amino acids

Note:

Depth, sequencing depth; Qual, quality of the called SP expressed in Phred score; Allele frequency, frequency of the alternative allele in each sample indicated in the "Sample" column.

tRNA-Ile, and *tRNA-Leu1*, all the other tRNA genes are organized in two big clusters: a 13-gene cluster positioned between *cox3* and *nd6*, and a 6-gene cluster between *nd6* and *nd4*. The assembly retrieved 6 out of 13 tRNAs from the first cluster (missing *tRNA-Gly*, *tRNA-*

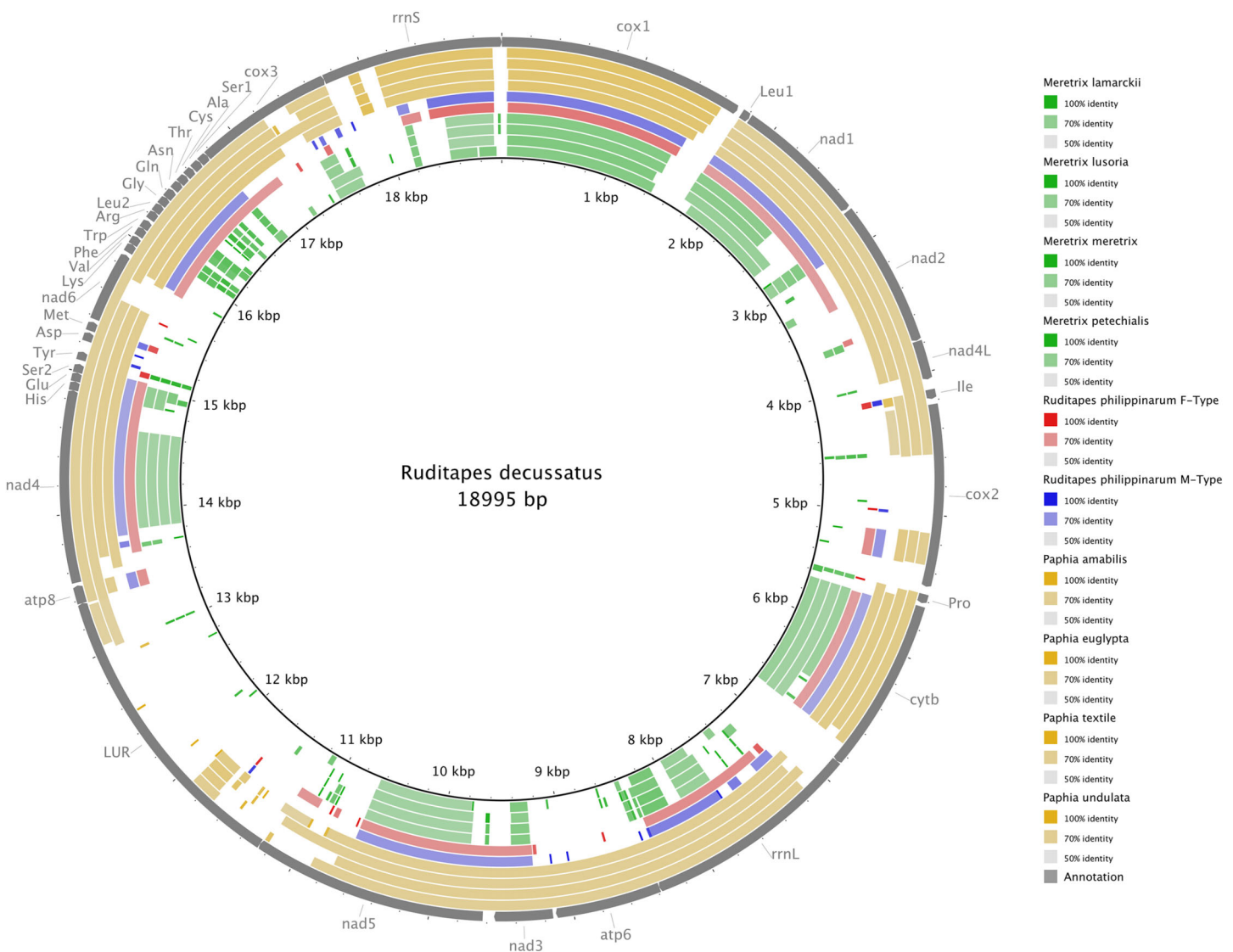


Figure 3 BLASTN comparison of *Ruditapes decussatus* and other Veneridae. *Ruditapes decussatus* mtDNA map (external gray circle), and BLASTN identity (colored inner circles) with complete mtDNAs of other 10 venerid species (see list in Table S3).

Glu, *tRNA-Asn*, *tRNA-Thr*, *tRNA-Cys*, *tRNA-Ala*, and *tRNA-Ser1*), and 4 out of 6 tRNAs from the second cluster (missing *tRNA-Met* and *tRNA-Asp*). All the tRNA genes not located in these two clusters (*tRNA-Pro*, *tRNA-Ile*, and *tRNA-Leu1*) were included in the contigs. The presence of a clear polyadenylation signal in four of the assembled contigs (see Table 1) seems to indicate the existence of multiple polycistronic transcripts. It is also noteworthy that poly-A sequences seem to be absent in contigs having tRNA or rRNA genes at one end (Contigs 2, 5, 8 and 9). This could be either an evidence supporting the “tRNA punctuation model” of RNA processing proposed by *Ojala, Montoya & Attardi (1981)* for human mitochondria, or a result of difficulties in sequencing/assembly of such regions. More analyses are required to address this point.

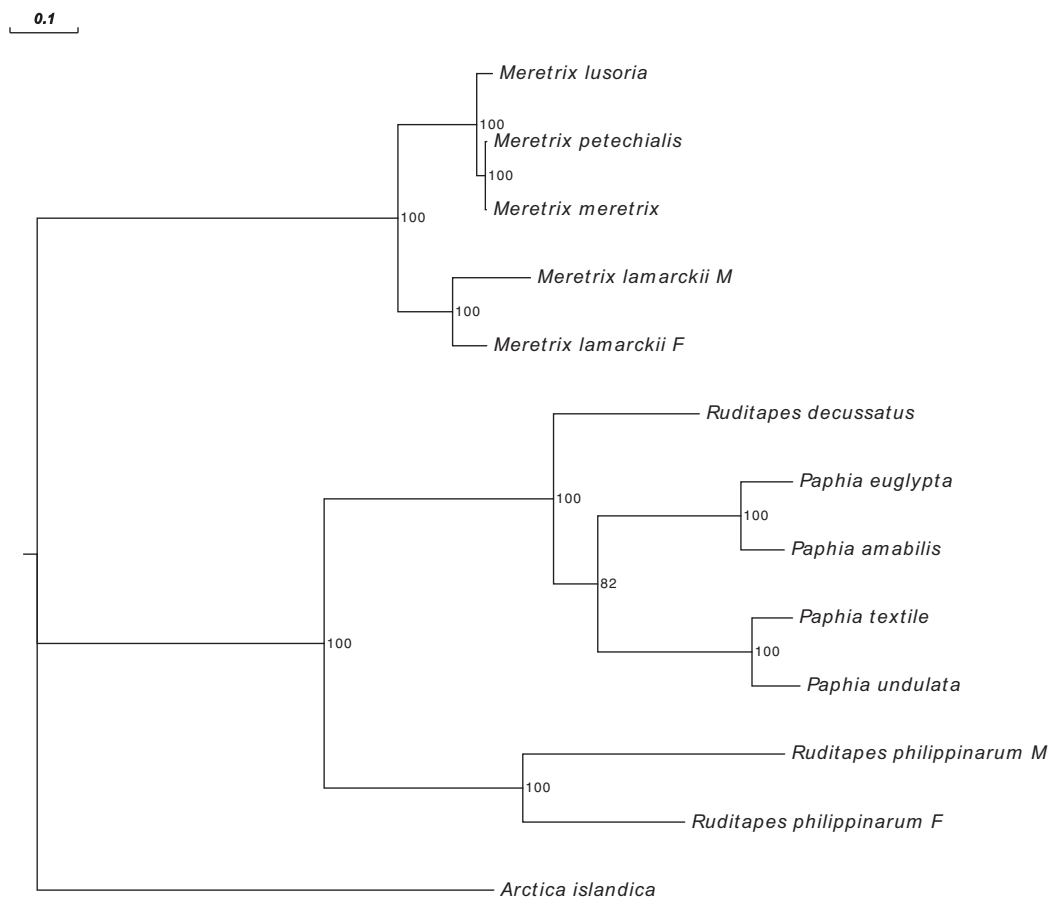


Figure 4 Maximum Likelihood (ML) tree of Veneridae obtained with all mitochondrial coding genes. ML tree obtained with the MitoPhast pipeline; the complete input and output of this analysis is available on figshare (<https://doi.org/10.6084/m9.figshare.4970762.v1>).

General features

The size of the fully Sanger-sequenced mitochondrial genome of *Ruditapes decussatus* (reference female F4) is of 18,995 bp, and it includes 13 protein-coding genes, 22 tRNAs and 2 rRNAs. Our data support the presence of the *atp8* gene in the mtDNA of *Ruditapes decussatus*; *atp8* has been reported as missing in several bivalve species, however, more accurate searches often led to the identification of the gene, so, in most cases, the alleged lack of *atp8* is likely ascribable to annotation inaccuracies due to the extreme variability and the small size of the gene (Breton, Stewart & Hoeh, 2010; Breton et al., 2014; Plazzi, Puccio & Passamonti, 2016).

The mitochondrial genome of *Ruditapes decussatus* shows a high content of A-T (63%), a common feature in bivalve mtDNAs; moreover, T is the most common nucleotide at the third codon base (64.6%). The most common codon is UUU (Phe), which is also the most commonly used in bivalves, as well as in other invertebrates (Passamonti et al., 2011).

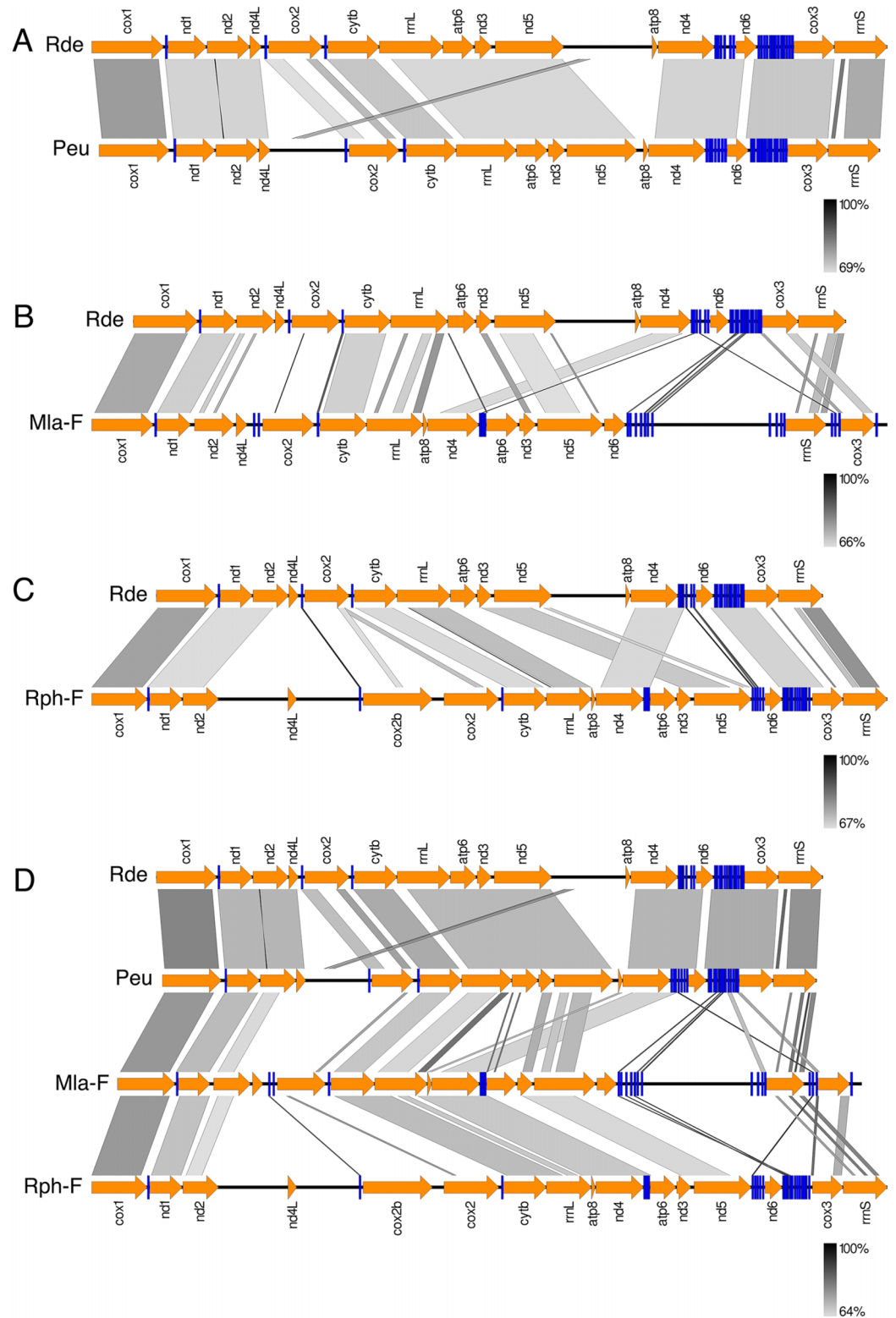


Figure 5 Comparison of gene order in venerid mtDNAs. Variation in gene order between *Ruditapes decussatus* and *P. euglypta* (A) *M. lamarckii* F-type (B) *Ruditapes philippinarum* F-type (C) and among all the four species (D).

Codon usage

As shown in Table 5, in 16 cases out of 20, the most frequently used codon does not correspond to the anticodon of the inferred tRNA. In other words, there is not a correspondence between the most abundant codons and the anticodons of the 22 mitochondrial tRNAs. According to the “wobble hypothesis”—first proposed by Crick (1966)—the conformation of the tRNA anticodon loop enables some flexibility at the first base of the anticodon, so a Watson–Crick type of base pairing in the third position of the codon is not strictly necessary. This allows an amino acid to be correctly incorporated by ribosomes even if the tRNA is not fully complementary to the codon; according to Crick, this explains the degeneracy of the genetic code. This feature is particularly interesting in the light of the debate about natural selection acting at synonymous sites: since the early 1980s, evidence of a correlation between synonymous codon usage and tRNA abundances started accumulating. According to these authors, synonymous codon usage is biased to match skews in tRNA abundance, as a result of selective pressure maximizing protein synthesis rates (reviewed in Chamary, Parmley & Hurst, 2006). Following this rationale, the results here reported and data from other marine bivalves and metazoans (Yu & Li, 2011; Passamonti et al., 2011) would suggest that in some mitochondrial genomes translation efficiency is not maximized, and this observation deserves further investigation.

Length and sequence polymorphism

The mtDNA of *Ruditapes decussatus* has a high proportion of URs mostly depending on the length of the LUR (Table 6); on average, bivalve mtDNAs have 1.7× the amount of URs in respect to other analyzed Metazoa (Ghiselli et al., 2013), and it is still unclear whether there is an accumulation of non-functional sequences in bivalve mtDNA due to genetic drift, or if such URs are maintained by natural selection because they contain—so far unknown—functional elements (see Milani et al., 2013, 2014b; Breton et al., 2014; Pozzi et al., 2017). The LUR of *Ruditapes decussatus* most likely includes the mitochondrial CR, as indicated by the presence of two motifs (Fig. 2C; Figs S2 and S3) similar to two regulatory elements identified in the sea urchin CR. These two motifs are the same identified in previous analyses on the clam *Ruditapes philippinarum* and the mussel *Musculista senhousia* (Ghiselli et al., 2013; Guerra, Ghiselli & Passamonti, 2014) so they are conserved across distant bivalve taxa, and the GO terms associated with such motifs are related to transcriptional control (Table S4). An interesting feature of *Ruditapes decussatus* LUR is its variable length (Table 6), most likely due to different repeat content. As a matter of fact, the very same repeat sequence was present in every LUR, and our data strongly suggest that LUR length variation is actually due to repeat CNV (see Supplemental Information files on figshare: <https://ndownloader.figshare.com/files/8387666> and <https://ndownloader.figshare.com/files/8360789>), as observed in other bivalve species (see Ghiselli et al., 2013; Guerra, Ghiselli & Passamonti, 2014). Tandem repeats have been also reported in the mitochondrial genomes of the bivalves *Acanthocardia tuberculata* (Dreyer & Steiner, 2006), *Placopecten magellanicus* (La Roche et al., 1990), *Moerella iridescens*, *Sanguinolaria olivacea*, *Semele scaba*, *Sinonovacula constricta*, *Solecurtus divaricatus*

(Yuan et al., 2012), *Ruditapes philippinarum* (Ghiselli et al., 2013), and *Musculista senhousia* (Guerra, Ghiselli & Passamonti, 2014). These repeats are believed to arise from duplications caused by replication slippage (Buroker et al., 1990; Hayasaka, Ishida & Horai, 1991; Broughton & Dowling, 1994). The tandem repeats found at the 3' end of *Ruditapes decussatus* LUR are predicted to form a secondary structure (see Fig. 2B and Supplemental Information files on figshare) composed by multiple stem-loops, which obviously increase in number with the increment of the number of tandem repeats. The effect, if any, of tandem repeats in mtDNA is unknown: since the repeats are almost always localized in proximity of the CR, they might interact with regulatory elements—or even contain some—influencing replication and/or transcription initiation, and such interactions might also be altered by the formation of secondary structures (Passamonti et al., 2011; Ghiselli et al., 2013; Guerra, Ghiselli & Passamonti, 2014).

We assessed the genetic variability of *Ruditapes decussatus* mtDNA using two different approaches: by SP calling in CDS (RNA-Seq data on 12 individuals), and by analysis of the LUR (Sanger sequencing of 14 individuals). The CR and its flanking regions are known to be hypervariable, so they are commonly used to assess polymorphism at low taxonomic levels. Our data strongly support a very low genetic variability: the number of SPs in CDS is 145, of which 103 are private of a single individual (mRDI01)—thus reducing the number to 42—while the number of variable sites in the analyzed LURs is 98 over 3,095 aligned positions. Considering the known variability of mtDNA in bivalves (Gissi, Iannelli & Pesole, 2008; Ghiselli et al., 2013; Breton et al., 2014; Plazzi, Puccio & Passamonti, 2016), this is a surprising result. Even more if we compare the results of the present work to a methodologically identical analysis performed on 12 *Ruditapes philippinarum* samples from the Pacific coast of USA, performed by Ghiselli et al. (2013): in that work, GATK yielded 194 SPs in the M-type mtDNA and 293 in the F-type. Strikingly, the 12 *Ruditapes philippinarum* samples analyzed were actually two families (6 siblings + 6 siblings). This means that randomly sampled individuals of *Ruditapes decussatus* used in this work showed a much lower mtDNA variability than *Ruditapes philippinarum* siblings. A previous analysis on the *cox1* gene of *Ruditapes decussatus* reported a nucleotide diversity (π) of 0.15 for a population from the Northern Adriatic Sea (Cordero, Peña & Saavedra, 2014). Another analysis on the same gene of *Ruditapes philippinarum* from the same range resulted in a $\pi = 0.25$ (Cordero et al., 2017), so *Ruditapes decussatus* has a lower nucleotide diversity at the *cox1* locus. The difference between the variability in mtDNA of *Ruditapes decussatus* that we are reporting here and that of *Ruditapes philippinarum* reported in Ghiselli et al. (2013) appears to be more marked. It is known that the genetic variability of *Ruditapes philippinarum* in the Adriatic Sea is lower than in populations from its native range in Asia (Cordero et al., 2017), probably because of the bottlenecks that this species had to go through during the multiple colonization events. The introduction in North America from Asia happened first (in the 1930s), and from there the Manila clam was introduced in Atlantic Europe (in the 1970s and 1980s), and lastly into the Adriatic Sea (1983 and 1984), and it is plausible that the genetic diversity decreased at each introduction event. Accordingly, Cordero et al. (2017) observed that *Ruditapes philippinarum* genetic variability in Europe is lower compared to that of the Pacific coast

of the USA, so the samples analyzed in *Ghiselli et al. (2013)* could have been more polymorphic than those analyzed in *Cordero, Peña & Saavedra (2014)*, thus explaining the more pronounced differences in genetic variability between the Manila clam and the European clam discussed above. In any case, all the available data point to a lower genetic diversity of *Ruditapes decussatus* mtDNA, and it would be interesting to know whether it is a cause or an effect of the ongoing replacement of *Ruditapes decussatus* with the invasive *Ruditapes philippinarum*. It will also be important to investigate genetic variability of the nuclear genes, especially after *Cordero, Peña & Saavedra (2014)* reported contrasting levels of differentiation between mitochondrial and nuclear markers.

With respect to SP effects, we found six indels in CDS, 2 of which do not cause frameshift, but a simple insertion/deletion of one amino acid (SP_1698, and SP_17619, see [Table 8](#)). Of the remaining four, SP_6364 and SP_10449 consist of a deletion and an insertion of a single T in two homopolymeric sequences (CTTTTTTTT and CTTTTTTT, respectively), raising the possibility of a sequencing error. In any case, the two SPs yield a shorter CDS (*cytb* and *nd5*, respectively), and are present at relatively low frequencies in the specimens carrying them, except for SP_6364 which has a frequency of 80% in fRDI04. The *cox3* gene shows three SPs: the first one, SP_17619, does not cause a frameshift, and results in the deletion of one alanine residue, and its frequency in mRDI01 is 97%. The second one, SP_17621, consists of a deletion of a G with respect to the reference sequence, which is the Sanger-sequenced mtDNA of sample F4; all the individuals analyzed with RNA-Seq carry this deletion except for mRDI01 which, at that position, has the same sequence of the reference mtDNA (reference-like allele frequency in mRDI01 = 99%). The third indel, SP_17624, consists of an insertion of two nucleotides, and its frequency in mRDI01 is 99%. So, basically, for *cox3* we have three types of sequences: (i) the Sanger-sequenced reference, which yields a 966 bp (321 aa) ORF; (ii) a sequence found in 11/12 of samples analyzed with RNA-Seq (except mRDI01) that carries a single-nucleotide deletion (SP_17621), and yields a 963 bp (320 aa) ORF; (iii) a sequence, private of mRDI01, which is obtained by combining SP_17624 and SP_17621 (both 99% of frequency, so most likely co-occurring), which produces a 963 bp (320 aa) ORF. Interestingly, the ORFs obtained from the sequences described in (ii) and (iii), are almost identical, namely the sequence obtained by RNA-seq in 11/12 samples and the sequence obtained by RNA-Seq in mRDI01 are basically the same, and differ from the Sanger-sequenced reference, yielding an amino acid sequence that differs in the last 35 residues (all data available in Supplemental Information files on figshare: <https://doi.org/10.6084/m9.figshare.4970762.v3>). Given this consistent difference between the sequence obtained by Sanger-sequencing of DNA, and those obtained by RNA-Seq, it is tempting to speculate that this difference might be caused by RNA editing, a mechanism observed in mtDNA of some animals (*Lavrov & Pett, 2016*), and recently reported to be common in cephalopods (*Liscovitch-Brauer et al., 2017*). Actually, *Liscovitch-Brauer et al. (2017)* reported only A-to-I editing, which is not the kind of change we are observing here, but other types of editing are known across eukaryotes (see *Gott & Emeson, 2000* for a review), and some others, still unknown, might exist as well. Post-transcriptional modifications (thus including RNA-editing) are still poorly understood mechanisms, but they appear to be responsible for

most of the mitochondrial gene expression regulation ([Scheibye-Alsing et al., 2007](#); [Scheffler, 2008](#); [Milani et al., 2014a](#)). What we propose here is a pure conjecture, but we think in the future it might be worthy to investigate mitochondrial transcriptomes looking for such kind of “unexpected” biological features.

Interestingly, in contrast with a low nucleotide variability along the entire mitochondrial genome, we observed a pretty high polymorphism in LUR length due to CNV of tandem repeats, and even a LUR length heteroplasmy: two females yielded two electrophoretic bands each (~2,100 and ~3,500 bp in F3; ~2,500 and ~3,500 bp in F17; see [Table 6](#)). A possible explanation is that the diversity (CNV) detected in the LURs could be recent: the accumulation of nucleotide variation at different sites along the mitochondrial genome needs time, while the kind structural variability we observed can be achieved in few generations (or even one) considering that replication slippage is common in repeat-rich regions.

Phylogenetic relationship with *Ruditapes philippinarum*

Despite *Ruditapes decussatus* and *Ruditapes philippinarum* being morphologically similar and being ascribed to the same genus, the results here reported clearly show that they are quite different both for mtDNA sequence ([Figs. 3 and 4](#)) and mtDNA gene arrangement ([Fig. 5](#)). This is an unusual finding, even among bivalves, which are known to be fast-evolving for these characters. This may point to the fact that these two species are less related than previously thought. Actually, this is not the first clue that *Ruditapes decussatus* and *Ruditapes philippinarum* are quite different genetically, as allozyme electrophoresis ([Passamonti, Mantovani & Scali, 1997, 1999](#)) and satellite DNA content ([Passamonti, Mantovani & Scali, 1998](#)) pointed out. More in-depth analyses are therefore needed to correctly trace the phylogenetic relationships of these two *Ruditapes* species, which may eventually end up in two different Genera. As shown in [Figs. 3–5](#), the Genus *Paphia* is the most similar to *Ruditapes decussatus*.

Presence/absence of DUI

We could not find evidence for sex-specific mtDNAs, typical of DUI. As stated in the Introduction, the search for DUI is not a straightforward process. HTS can help thanks to a much deeper sequencing coverage (in respect to the cloning-and-Sanger-sequencing approach), and because it overcomes the problem of primer specificity, a limitation of the classical approach. One possible concern about using HTS approaches based on short reads in presence of DUI is about the ability of softwares to detect divergent reads and assembly them correctly. More specifically, one could ask what is the divergence threshold under which the assemblers are not able to partition the contigs into two sex-linked groups. We do not know such a threshold, but we used different assembly strategies trying to retrieve sex-specific mtDNA sequences from our data. Other than the approach reported in Materials and Methods (which is the one that produced the data reported here), we tried other techniques. After identifying reads that blasted to bivalve mitochondrial sequences present in GenBank and discarding all the other reads, we generated A5+CAP3 assemblies: (i) for each of the samples (obtaining 12 separate assemblies), and (ii) pooling the six males together and the six females together, and

assembling the two sex-specific pools. Both these approaches did not show evidence of sex-specific mtDNAs. Then we took the assembly obtained from the females and removed the reads from each of the samples that mapped (<8 mismatches) to these sequences. We then used the remaining reads as A5 input. The program could not assemble anything. Lastly, we tried the software MetaVelvet ([Namiki et al., 2012](#))—that assembles metagenomes—on all the reads matching bivalve mtDNAs, and only one genome was produced. After all these alternative approaches failed to find two sex-linked mtDNAs, we decided to proceed with the assembly as indicated in Materials and Methods, because it was the technique that yielded the best quality contigs, most likely because using the reads from all 12 the individuals granted a higher coverage of the mtDNA. Given these results, we can propose three different explanations.

1. *Ruditapes decussatus* is characterized by SMI of mitochondria, so a male-transmitted mtDNA is not present in this species.
2. The divergence between the two sex-specific mtDNAs is too low to be detected. This could be the outcome of two different situations.
 - a) DUI is very young in this species, so the two sex-linked mtDNAs did not have the time to diverge.
 - b) A role-reversal event occurred recently. Role reversal (a.k.a. “route reversal” or “masculinization”) is a process—observed so far only in species of the *Mytilus* complex—by which F-type genomes invades the male germ line becoming sperm-transmitted, thus turning into M-type mtDNAs ([Hoeh et al., 1997](#)). This event actually resets to zero the divergence between F- and M-type, although substantial differences in the control regions were reported between the original F-type and the “masculinized” one (see [Zouros, 2013](#) for a thorough review). The hypothesis that role reversal could have occurred multiple times in the evolutionary history of bivalves and could have led to the complete replacement of M or F mtDNAs in several species was proposed by [Hoeh et al. \(1997\)](#) to explain the scattered phylogenetic distribution of DUI across Bivalvia. Indeed, according to the hypothesis of a single origin, DUI arose >400 Mya, approximately at the origin of Autolamellibranchia, but, as said, such hypothesis requires the assumption of multiple role-reversal and/or DUI loss events in several branches of the bivalve tree (see [Zouros, 2013](#) for a detailed discussion). Recently, a multiple origin of DUI was proposed ([Milani et al., 2013, 2014b](#); [Milani, Ghiselli & Passamonti, 2016](#); [Mitchell et al., 2016](#)), and in such case there would be no need of multiple role-reversal events to explain its phylogeny. In our opinion, until further evidence will be provided, role-reversal should not be considered a rule, but rather an exception. Of course, we cannot rule out that a masculinization event might have occurred in *Ruditapes decussatus*, so this hypothesis must be taken into consideration.
3. In our data, even if there is no clear evidence of a male-specific mtDNA, a male sample (mRDI01) clearly stood out from the others, both males and females (see [Table 7](#)). Overall, the divergence between mRDI01 and the other 11 samples calculated

considering its private SPs is of 151 sites over 18,995 bp (considering the whole mtDNA), and of 103 sites over 14,920 bp (considering only CDS). In both cases the divergence is very low (0.8% and 0.7%, respectively), which explains why the mtDNA of mRDI01, although different, was not assembled as a separate genome. We have no sufficient data to evaluate if such divergence is normal within *Ruditapes decussatus* populations, but considering the variability usually observed in bivalves, we find the difference unsurprising. On the contrary, the lack of variability among the other 11 samples is remarkable. For these reasons, we are inclined to believe that mRDI01 divergence is compatible with hypotheses (1) and (2). That said, there still could be a third, quite conjectural, hypothesis by which these data might indicate an incipient DUI, not yet fixed in the population.

All in all, we have a preference for the first explanation, but the present data are not sufficient to exclude the others, and a more thorough investigation is necessary to assess this point.

Up to now DUI was identified in only three Veneridae species: *Cyclina sinensis*, *Ruditapes philippinarum*, and *Meretrix lamarckii* (Gusman *et al.*, 2016). The status of *Paphia* is still unknown, and in future works it would be interesting to investigate more Heterodonta species to understand better the distribution of DUI in this derived group of bivalves.

ACKNOWLEDGEMENTS

We would like to thank Edoardo Turolla (Istituto Delta Ecologia Applicata, Ferrara, Italy) for providing the specimens, and Massimo Milan for bibliographic suggestions. We also gratefully thank the Editor Tim Collins, and the reviewers Carlos Saavedra, Shallee Page, and one anonymous colleague for their comments and suggestions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by the Italian Ministry of Education, University and Research (MIUR) FIR Programme no. RBFR13T97A funded to FG, MIUR SIR Programme no. RBSI14G0P5 funded to LM, Zumberge Foundation to SVN, and by the Canziani bequest funded to MP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Italian Ministry of Education, University and Research (MIUR) FIR: RBFR13T97A.

MIUR SIR: RBSI14G0P5.

Zumberge Foundation.

Canziani bequest.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Fabrizio Ghiselli conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Liliana Milani conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, reviewed drafts of the paper.
- Mariangela Iannello performed the experiments, analyzed the data, prepared figures and/or tables, reviewed drafts of the paper.
- Emanuele Procopio performed the experiments, analyzed the data.
- Peter L. Chang analyzed the data.
- Sergey V. Nuzhdin conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Marco Passamonti conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

GenBank accession numbers [MF055702](#) to [MF055714](#), and [KP089983](#).

GenBank BioProject [PRJNA170478](#).

Data Availability

The following information was supplied regarding data availability:

Ghiselli, Fabrizio; Milani, Liliana; Iannello, Mariangela; Procopio, Emanuele; L. Chang, Peter; Nuzhdin, Sergey; Passamonti, Marco (2017): The Complete Mitochondrial Genome of the Grooved Carpet Shell, *Ruditapes decussatus* (Bivalvia, Veneridae). figshare. <https://doi.org/10.6084/m9.figshare.4970762.v3>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.3692#supplemental-information>.

REFERENCES

- Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**(1):402 DOI [10.1186/1471-2164-12-402](#).
- Arias-Pérez A, Cordero D, Borrell Y, Sánchez JA, Blanco G, Freire R, Insua A, Saavedra C. 2016. Assessing the geographic scale of genetic population management with microsatellites and introns in the clam *Ruditapes decussatus*. *Ecology and Evolution* **6**(10):3380–3404 DOI [10.1002/ece3.2052](#).
- Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. 2003. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research* **31**(1):400–402 DOI [10.1093/nar/gkg030](#).

- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* 37:W202–W208 DOI 10.1093/nar/gkp335.
- Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsche G, Pütz J, Middendorf M, Stadler PF. 2013. MITOS: improved *de novo* metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* 69(2):313–319 DOI 10.1016/j.ympev.2012.08.023.
- Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M. 2014. A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. *Trends in Genetics* 30(12):555–564 DOI 10.1016/j.tig.2014.09.002.
- 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. *Marine Genomics* 3(1):11–18 DOI 10.1016/j.margen.2010.01.001.
- Breton S, Stewart DT, Shepardson S, Trdan RJ, Bogan AE, Chapman EG, Ruminas AJ, Piontkivska H, Hoeh WR. 2011. Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? *Molecular Biology and Evolution* 28(5):1645–1659 DOI 10.1093/molbev/msq345.
- Broughton RE, Dowling TE. 1994. Length variation in mitochondrial DNA of the minnow *Cyprinella spiloptera*. *Genetics* 138:179–190.
- Buroker NE, Brown JR, Gilbert TA, O’Hara PJ, Beckenbach AT, Thomas WK, Smith MJ. 1990. Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. *Genetics* 124:157–163.
- Buske FA, Bodén M, Bauer DC, Bailey TL. 2010. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* 26(7):860–866 DOI 10.1093/bioinformatics/btq049.
- Cao L, Kenchington E, Zouros E, Rodakis GC. 2004. Evidence that the large noncoding sequence is the main control region of maternally and paternally transmitted mitochondrial genomes of the marine mussel (*Mytilus* spp.). *Genetics* 167(2):835–850 DOI 10.1534/genetics.103.026187.
- Capella-Gutiérrez S, Silla-Martinez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973 DOI 10.1093/bioinformatics/btp348.
- Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics* 7(2):98–108 DOI 10.1038/nrg1770.
- Cordero D, Delgado M, Liu B, Ruesink J, Saavedra C. 2017. Population genetics of the Manila clam (*Ruditapes philippinarum*) introduced in North America and Europe. *Scientific Reports* 7:39745 DOI 10.1038/srep39745.
- Cordero D, Peña JB, Saavedra C. 2014. Phylogeographic analysis of introns and mitochondrial DNA in the clam *Ruditapes decussatus* uncovers the effects of Pleistocene glaciations and endogenous barriers to gene flow. *Molecular Phylogenetics and Evolution* 71:274–287 DOI 10.1016/j.ympev.2013.11.003.
- Crick FHC. 1966. Codon—anticodon pairing: the wobble hypothesis. *Journal of Molecular Biology* 19(2):548–555 DOI 10.1016/s0022-2836(66)80022-0.
- Darty K, Denise A, Ponty Y. 2009. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25(15):1974–1975 DOI 10.1093/bioinformatics/btp250.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43(5):491–498 DOI 10.1038/ng.806.

- de Sousa JT, Milan M, Bargelloni L, Pauletto M, Matias D, Joaquim S, Matias AM, Quillien V, Leitão A, Huvet A. 2014. A microarray-based analysis of gametogenesis in two Portuguese populations of the European clam *Ruditapes decussatus*. *PLOS ONE* 9:e92202 DOI 10.1371/journal.pone.0092202.
- Diz AP, Dudley E, Skibinski DOF. 2012. Identification and characterisation of highly expressed proteins in sperm cells of the marine mussel *Mytilus edulis*. *Proteomics* 12(12):1949–1956 DOI 10.1002/pmic.201100500.
- Dreyer H, Steiner G. 2006. The complete sequences and gene organisation of the mitochondrial genomes of the heterodont bivalves *Acanthocardia tuberculata* and *Hiattella arctica*—and the first record for a putative Atpase subunit 8 gene in marine bivalves. *Frontiers in Zoology* 3:13.
- Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44(D1):D279–D285 DOI 10.1093/nar/gkv1344.
- Ghiselli F, Milani L, Chang PL, Hedgecock D, Davis JP, Nuzhdin SV, Passamonti M. 2012. *De Novo* assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. *Molecular Biology and Evolution* 29(2):771–786 DOI 10.1093/molbev/msr248.
- 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 Biology and Evolution* 5(8):1535–1554 DOI 10.1093/gbe/evt112.
- Ghiselli F, Milani L, Passamonti M. 2011. Strict sex-specific mtDNA segregation in the germ line of the DUI species *Venerupis philippinarum* (Bivalvia: Veneridae). *Molecular Biology and Evolution* 28(2):949–961 DOI 10.1093/molbev/msq271.
- Gissi C, Iannelli F, Pesole G. 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity* 101(4):301–320 DOI 10.1038/hdy.2008.62.
- Gosling EM. 2003. *Bivalve Molluscs: Biology, Ecology and Culture*. Hoboken: Wiley Online Library.
- Gott JM, Emeson RB. 2000. Functions and mechanisms of RNA editing. *Annual Review of Genetics* 34(1):499–531 DOI 10.1146/annurev.genet.34.1.499.
- Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Research* 36:W70–W74 DOI 10.1093/nar/gkn188.
- Guerra D, Ghiselli F, Passamonti M. 2014. The Largest Unassigned Regions of the male- and female-transmitted mitochondrial DNAs in *Musculista senhousia* (Bivalvia Mytilidae). *Gene* 536(2):316–325 DOI 10.1016/j.gene.2013.12.005.
- 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 DOI 10.7717/peerj.2760.
- Hayasaka K, Ishida T, Horai S. 1991. Heteroplasmy and polymorphism in the major noncoding region of mitochondrial DNA in Japanese monkeys: association with tandemly repeated sequences. *Molecular Biology and Evolution* 8:399–415 DOI 10.1093/oxfordjournals.molbev.a040660.
- He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S. 2016. Evolvview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research* 44(W1):W236–W241 DOI 10.1093/nar/gkw370.

- Hoeh WR, Stewart DT, Saavedra C, Sutherland BW, Zouros E. 1997. Phylogenetic evidence for role-reversals of gender-associated mitochondrial DNA in *Mytilus* (Bivalvia: Mytilidae). *Molecular Biology and Evolution* 14(9):959–967 DOI 10.1093/oxfordjournals.molbev.a025839.
- Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Research* 9(9):868–877 DOI 10.1101/gr.9.9.868.
- Jones P, Binns D, Chang H-YY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, 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 DOI 10.1093/bioinformatics/btu031.
- Ju YS, Kim J-I, Kim S, Hong D, Park H, Shin J-Y, Lee S, Lee W-C, Kim S, Yu S-B, Park SS, Seo S-H, Yun J-Y, Kim H-J, Lee D-S, Yavartanoo M, Kang HP, Gokcumen O, Govindaraju DR, Jung JH, Chong H, Yang K-S, Kim H, Lee C, Seo J-S. 2011. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nature Genetics* 43(8):745–752 DOI 10.1038/ng.872.
- King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Science International: Genetics* 12:128–135 DOI 10.1016/j.fsigen.2014.06.001.
- La Roche J, Snyder M, Cook DI, Fuller K, Zouros E. 1990. Molecular characterization of a repeat element causing large-scale size variation in the mitochondrial DNA of the sea scallop *Placcopecten magellanicus*. *Molecular Biology and Evolution* 7(1):45–64 DOI 10.1093/oxfordjournals.molbev.a040586.
- Laslett D, Canback B. 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* 24(2):172–175 DOI 10.1093/bioinformatics/btm573.
- Lavrov DV, Pett W. 2016. Animal Mitochondrial DNA as We Do Not Know It: mt-Genome Organization and Evolution in Nonbilaterian Lineages. *Genome Biology and Evolution* 8(9):2896–2913 DOI 10.1093/gbe/evw195.
- Leite RB, Milan M, Coppe A, Bortoluzzi S, dos Anjos A, Reinhardt R, Saavedra C, Patarnello T, Cancela ML, Bargelloni L. 2013. mRNA-Seq and microarray development for the Grooved Carpet shell clam, *Ruditapes decussatus*: a functional approach to unravel host-parasite interaction. *BMC Genomics* 14(1):741 DOI 10.1186/1471-2164-14-741.
- Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJC, Eisenberg E. 2017. Trade-off between transcriptome plasticity and genome evolution in cephalopods. *Cell* 169(2):191–202.e11 DOI 10.1016/j.cell.2017.03.025.
- Lubośny M, Przyłucka A, Sańko TJ, Śmietanka B, Rosenfeld S, Burzyński A. 2017/2. Next generation sequencing of gonadal transcriptome suggests standard maternal inheritance of mitochondrial DNA in *Eurhomalea rufa* (Veneridae). *Marine Genomics* 31:21–23 DOI 10.1016/j.margen.2016.11.002.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20(9):1297–1303 DOI 10.1101/gr.107524.110.
- Milani L, Ghiselli F. 2015. Mitochondrial activity in gametes and transmission of viable mtDNA. *Biology Direct* 10:22 DOI 10.1186/s13062-015-0057-6.
- Milani L, Ghiselli F, Guerra D, Breton S, Passamonti M. 2013. A comparative analysis of mitochondrial ORFans: new clues on their origin and role in species with doubly uniparental

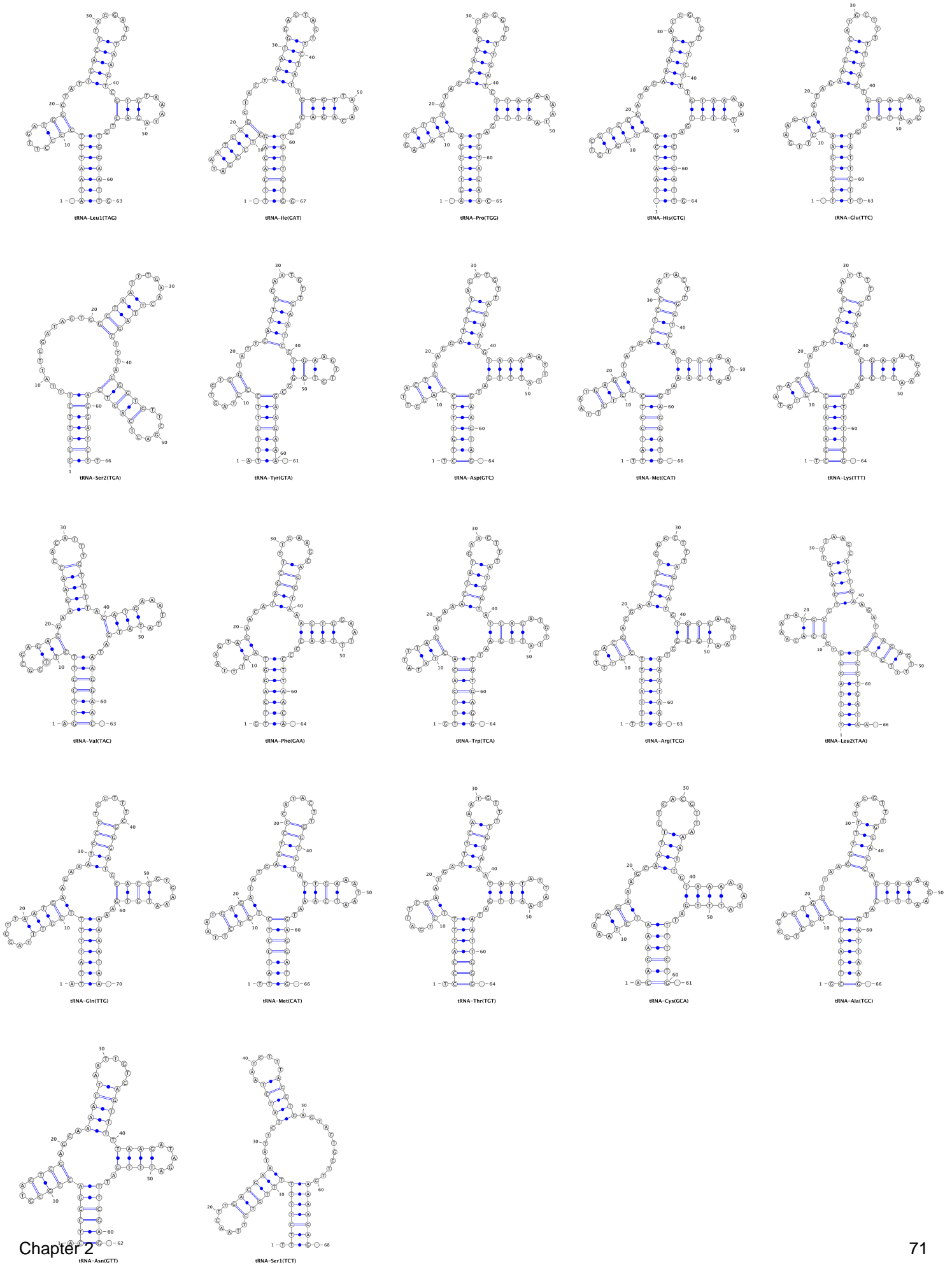
inheritance of mitochondria. *Genome Biology and Evolution* 5(7):1408–1434
DOI 10.1093/gbe/evt101.

- Milani L, Ghiselli F, Iannello M, Passamonti M. 2014a.** Evidence for somatic transcription of male-transmitted mitochondrial genome in the DUI species *Ruditapes philippinarum* (Bivalvia: Veneridae). *Current Genetics* 60(3):163–173 DOI 10.1007/s00294-014-0420-7.
- Milani L, Ghiselli F, Maurizii MG, Nuzhdin SV, Passamonti M. 2014b.** Paternally transmitted mitochondria express a new gene of potential viral origin. *Genome Biology and Evolution* 6(2):391–405 DOI 10.1093/gbe/evu021.
- Milani L, Ghiselli F, Passamonti M. 2016.** Mitochondrial selfish elements and the evolution of biological novelties. *Current Zoology* 62(6):687–697 DOI 10.1093/cz/zow044.
- Milbury CA, Lee JC, Cannone JJ, Gaffney PM, Gutell RR. 2010.** Fragmentation of the large subunit ribosomal RNA gene in oyster mitochondrial genomes. *BMC Genomics* 11(1):485 DOI 10.1186/1471-2164-11-485.
- Mitchell A, Guerra D, Stewart D, Breton S. 2016.** *In silico* analyses of mitochondrial ORFans in freshwater mussels (Bivalvia: Unionoida) provide a framework for future studies of their origin and function. *BMC Genomics* 17(1):597 DOI 10.1186/s12864-016-2986-6.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008.** Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5(7):621–628 DOI 10.1038/nmeth.1226.
- Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012.** MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research* 40(20):e155 DOI 10.1093/nar/gks678.
- Ojala D, Montoya J, Attardi G. 1981.** tRNA punctuation model of RNA processing in human mitochondria. *Nature* 290(5806):470–474 DOI 10.1038/290470a0.
- Passamonti M, Ghiselli F. 2009.** Doubly uniparental inheritance: two mitochondrial genomes, one precious model for organelle DNA inheritance and evolution. *DNA and Cell Biology* 28(2):79–89 DOI 10.1089/dna.2008.0807.
- Passamonti M, Mantovani B, Scali V. 1997.** Allozymic characterization and genetic relationships among four species of Tapetinae (Bivalvia, Veneridae). *Italian Journal of Zoology* 64:117–124 DOI 10.1080/11250009709356183.
- Passamonti M, Mantovani B, Scali V. 1998.** Characterization of a highly repeated DNA family in tapetinae species (mollusca bivalvia: veneridae). *Zoological Science* 15(4):599–605 DOI 10.2108/zsj.15.599.
- Passamonti M, Mantovani B, Scali V. 1999.** Allozymic analysis of some Mediterranean Veneridae (Mollusca: Bivalvia): preliminary notes on taxonomy and systematics of the family. *Journal of the Marine Biological Association of the UK* 79(5):899–906 DOI 10.1017/s0025315498001064.
- Passamonti M, Ricci A, Milani L, Ghiselli F. 2011.** Mitochondrial genomes and Doubly Uniparental Inheritance: new insights from *Musculista senhousia* sex-linked mitochondrial DNAs (Bivalvia Mytilidae). *BMC Genomics* 12(1):442 DOI 10.1186/1471-2164-12-442.
- Pesole G, Allen JF, Lane N, Martin W, Rand DM, Schatz G, Saccone C. 2012.** The neglected genome. *EMBO Reports* 13:473–474.
- Plazzi F, Puccio G, Passamonti M. 2016.** Comparative large-scale mitogenomics evidences clade-specific evolutionary trends in mitochondrial DNAs of Bivalvia. *Genome Biology and Evolution* 8(8):2544–2564 DOI 10.1093/gbe/evw187.
- Pozzi A, Plazzi F, Milani L, Ghiselli F, Passamonti M. 2017.** SmithRNAs: could mitochondria “bend” nuclear regulation? *Molecular Biology and Evolution* 34(8):1960–1973 DOI 10.1093/molbev/msx140.

- Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* 132:365–386 DOI 10.1385/1-59259-192-2:365.
- Scheffler IE. 2008. *Mitochondria*. Hoboken, NJ: Wiley.
- Scheibye-Alsing K, Cirera S, Gilchrist MJ, Fredholm M, Gorodkin J. 2007. EST analysis on pig mitochondria reveal novel expression differences between developmental and adult tissues. *BMC Genomics* 8(1):367 DOI 10.1186/1471-2164-8-367.
- Serb JM, Lydeard C. 2003. Complete mtDNA sequence of the North American freshwater mussel, *Lampsilis ornata* (Unionidae): an examination of the evolution and phylogenetic utility of mitochondrial genome organization in Bivalvia (Mollusca). *Molecular Biology and Evolution* 20(11):1854–1866 DOI 10.1093/molbev/msg218.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7(1):539 DOI 10.1038/msb.2011.75.
- Skibinski DO, Gallagher C, Beynon CM. 1994a. Mitochondrial DNA inheritance. *Nature* 368:817–818.
- Skibinski DO, Gallagher C, Beynon CM. 1994b. Sex-limited mitochondrial DNA transmission in the marine mussel *Mytilus edulis*. *Genetics* 138:801–809.
- Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. *Briefings in Functional Genomics* 12(5):454–456 DOI 10.1093/bfpg/els066.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313 DOI 10.1093/bioinformatics/btu033.
- Stothard P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* 28:1102–1104.
- Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* 27(7):1009–1010 DOI 10.1093/bioinformatics/btr039.
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* 30(12):2725–2729 DOI 10.1093/molbev/mst197.
- Tan MH, Gan HM, Schultz MB, Austin CM. 2015. MitoPhAST, a new automated mitogenomic phylogeny tool in the post-genomic era with a case study of 89 decapod mitogenomes including eight new freshwater crayfish mitogenomes. *Molecular Phylogenetics and Evolution* 85:180–188 DOI 10.1016/j.ympev.2015.02.009.
- Theologidis I, Fodelianakis S, Gaspar MB, Zouros E. 2008. Doubly uniparental inheritance (DUI) of mitochondrial DNA in *Donax trunculus* (Bivalvia: Donacidae) and the problem of its sporadic detection in Bivalvia. *Evolution; International Journal of Organic Evolution* 62(4):959–970 DOI 10.1111/j.1558-5646.2008.00329.x.
- Tritt A, Eisen JA, Facciotti MT, Darling AE. 2012. An integrated pipeline for *de novo* assembly of microbial genomes. *PLOS ONE* 7(9):e42304 DOI 10.1371/journal.pone.0042304.
- Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. 2005. Database resources of the National Center for Biotechnology. *Nucleic Acid Research* 33:D39–D45 DOI 10.1093/nar/gki062.
- Yu H, Li Q. 2011. Mutation and selection on the wobble nucleotide in tRNA anticodons in marine bivalve mitochondrial genomes. *PLOS ONE* 6(1):e16147 DOI 10.1371/journal.pone.0016147.

- Yuan Y, Li Q, Yu H, Kong L. 2012.** The complete mitochondrial genomes of six heterodont bivalves (Tellinoidea and Solenoidea): variable gene arrangements and phylogenetic implications. *PLOS ONE* 7(2):e32353 DOI [10.1371/journal.pone.0032353](https://doi.org/10.1371/journal.pone.0032353).
- Yuan S, Xia Y, Zheng Y, Zeng X. 2016.** Next-generation sequencing of mixed genomic DNA allows efficient assembly of rearranged mitochondrial genomes in *Amolops chunganensis* and *Quasipaa boulengeri*. *PeerJ* 4:e2786 DOI [10.7717/peerj.2786](https://doi.org/10.7717/peerj.2786).
- Zouros E. 2013.** Biparental inheritance through uniparental transmission: the doubly uniparental inheritance (DUI) of mitochondrial DNA. *Evolutionary Biology* 40(1):1–31 DOI [10.1007/s11692-012-9195-2](https://doi.org/10.1007/s11692-012-9195-2).
- Zouros E, Ball AO, Saavedra C, Freeman KR. 1994a.** Mitochondrial DNA inheritance. *Nature* 368:818.
- Zouros E, Oberhauser Ball A, Saavedra C, Freeman KR. 1994b.** An unusual type of mitochondrial DNA inheritance in the blue mussel *Mytilus*. *Proceedings of the National Academy of Sciences of USA* 91(16):7463–7467 DOI [10.1073/pnas.91.16.7463](https://doi.org/10.1073/pnas.91.16.7463).

Supplementary Figure 1- mitochondrial tRNA structures of *Ruditapes decussatus*



Name	Strand	Start	<i>P</i> -value		Sites	
PhUn	+	695	2.72e-19	AAAACTC AAA	GAGGGGGGGGAAAGGGGGGGGTCCCAAA	AAATTCCTTT
MeLa M	+	539	1.87e-18	GGAGCAG GCT	TGGGGGGGGGAAAAGGGGGGGGGAAAA	AACAATAARA
RuPhM	+	3016	1.19e-17	GGAAAA TAC	TGTGGGGGGGAAAGGGGGGGGTCTCCAGA	AACTCCTTC
RuPhF	+	3428	2.98e-17	GGACGTT GCT	ATGGGGGGGGAAAGGGGGGGGTCTCCAGA	AACTCCTTC
MeLu	+	1230	5.38e-17	CAAGCTT AAA	TGGGGGGGGGAAAATGGGGGGGGGAAAA	AGAAAAATA A
SeSc	+	82	9.47e-17	AGAAATG TGC	GTGGGGGGGGAAAGGGGGGGGGCCGTTAG	GCCGGAGAG G
MeMe	+	1240	1.88e-16	GCAAGCT TAA	GTGGGGGGGGAAAATGGGGGGGGCAGAAA	AATAGAAAA G
MePe	+	1240	4.68e-16	CAAGCTT AAG	TGGGGGGGGGAAAATGGGGGGGGGGGGG	GGCAGAAA A
MeLaF	+	558	4.68e-16	GGAGCAA GCT	TGGGGGGGGGAAAAGGGGGGGGGAAAA	GACGATTAAA
SoDi	+	34	1.26e-15	CTCCTCT GT	AATGGGGGGGGTAGGGGGGGGTGCAAAAA	TGGAAGGAA A
PhTe	+	722	2.56e-15	TGTGTAGT TT	TTTGGGGGGGAAAGGGGGGGCACTAAAA	AAATTCCTTT
StPu	+	72	4.05e-15	CTCTTTG CA	TATGGGGGGGGGGGGGGGGGACTCTAAA	TAATATATA
PhEu	+	969	1.09e-14	AAGACGT TTC	TATGGGGGGGAAAGGGGGGGTCCCTAAAA	AAACTCCTTT
PhAm	+	2305	1.09e-14	TGTTTGTT TC	TATGGGGGGGAAAGGGGGGGCCCTTAAAA	AAACTCCTTT
MeLy	+	1159	2.27e-14	TGTGTATC TA	ATGGGGGGGGTAAGGGGGGGGTATAAAG	TTGGTTATAT
RuDe	+	776	1.46e-13	ACAAACT ACC	GTAAGGGGGGATAAGGGGGGTTCGCAAAA	AAACTCCTTT
SoDp	+	153	2.90e-12	AGTGGTG TTG	GATGTGGGGGGGTAGGGGGGGACCGAGCT	TCCGAAGGAA
SiCo	+	730	5.37e-12	ATATAAA ATG	GAGCAGGGGGGAGGGGGAGGAGGAAAG	GAGCCCCCTT
FuMu	+	1386	7.13e-11	GGTATAA AAG	GTAAGGGGGGAGAGGGGAAGGAGATTAA	ACCTCCAAA
C0An	+	552	7.60e-11	GTGAAAA AAA	GGGTGGGAGTAAAGGGGGGTGAGGGCAGG	CAAGGCCCT
AcTu	+	1007	2.18e-10	CCCCCTT TC	TTTAAAGGGGGGAGGGGGGGTTCATCCA	CCCCCAAAG
MoIr	+	94	3.50e-10	AAGGAAA AAA	AGGGGGGGGGGTGTGTGTTATCCGGAGT	AGAAGAGGA C
MoIr	+	94	3.50e-10	AAGGAAA AAA	AGGGGGGGGGGTGTGTGTTATCCGGAGT	AGAAGAGGA C
NuOl	+	591	3.70e-10	CCCCTATG GG	GGGAGGGGGGCCCGGGAAGGGCCCCC	TCCCCCATA
LuRh	+	331	8.60e-10	TATTGAG GCA	GGGAGGGAGGAGATTAGGGATCCGGGAG	GGATCAAAAA

Name	Strand	Start	<i>p</i> -value	Sites		
MeLy	+	2152	2.76e-18	ATACATTATT	GGGGGGAGGGGGTCTAAGGGGGAGGG	GGGGTTGCT
MePe	+	1717	1.31e-16	TCTTTTGTTA	GGGGAGGGGGGTTTAAGGGGGAGGG	GAATGTGATA
MeMe	+	1708	1.31e-16	TCTTTTGTTA	GGGGAGGGGGGTTTAAGGGGGAGGG	GAATATGATA
MeLu	+	1701	1.31e-16	TCTTTAATTA	GGGGAGGGGGGTTTAAGGGGGAGGG	GAATGTGATT
SiCo	+	1062	1.11e-14	TAGGTAAATT	GGGGGAAGGGGGTCCGGGGGAAGGG	GGGTAGCGTG
RuPhM	+	3514	2.27e-13	ATTA AAAATGA	AAGAGGGGGGGGTCCTAGGGGGGGGG	AGATTTTAGA
RuPhF	+	3929	5.15e-12	TGGTATATAA	GATTGGGGGGGGGTTTGGGGGGGTTA	TTTAGAATTA
PhUn	+	495	1.82e-11	TGGGGTAAGG	GGTTGTAAGTGTAGTGTGGGCCGGGGGG	ACCACACTAT
PhEu	+	740	1.82e-11	TAGGATAAAG	GATTGTAAGTGCAGTATAGGCCGGGGGG	ACTATACCGT
MeLaM	+	1092	1.82e-11	TGTGTGTGTT	TGGGGGAGAGGGGGAATGAGGGGTGG	CTATGGTTAG
MeLaF	+	1716	2.02e-11	ATAGAGGCTG	GGGGGGTGGGGTTTGTTCAGTTAGGG	TATGGAAGTA
PhAm	+	2078	4.08e-11	TAGGATAAAG	GATTGTAAGTGCAGTGTGGGCCGGGGGG	ACCACATTGT
SeSc	+	833	5.48e-11	CAGGATTTC	TGGGGGAGAGATTCTGTGGGGGGGGG	GTAACGACA
PhTe	+	528	8.88e-11	TGGGAAAAAT	AGTTGTAAGTGCAGTATGGGCCGGGGGG	ACTATATTGT
FuMu	-	1705	1.56e-10	GTATCCGAA	GGGGGAGGGGGGAATTGGGAGGTTTA	ATCTCCCTTC
SoDi	+	694	1.71e-10	ATTGATGTAG	GGTGGATAGTGGGGATGTGCTGGTGTTGG	GATAAATTAT
C0An	+	804	1.01e-09	GTTTAGTTGA	GTTGGGATGGGGTAAGAGGAAGAAGGG	GAAAGCAAGA
StPu	-	9	1.67e-09	CCCGAACGAC	GAAGGAGAGAGGGGTGGGAGCTGTGATGG	GGATTCTT
LuRh	+	747	2.13e-09	TTATGATAGA	GGGGCGTGGGGCTTATCCGGGTTATTG	TTGATAGGGT
MoIr	+	1344	6.40e-08	TGTTAATTTT	GGTTTGAATGGTATGGGGGGAGTGTGTG	TAATTGTTTG
RuDe	-	1100	6.84e-08	CTCCATTAC	GTTACGGAGGGGCAAGTGTGGGGTTTA	CTGGCGCCCG
NuOl	+	127	1.01e-07	TAAATTTAAT	TTGGCCGGGAGGAGAAAAAGGGTAAGGC	AAACTAAAGT
SoDp	+	828	1.58e-07	TACTTGTGCT	ATGGTTTGGTGGTACTTGGTAGGAAGGG	ATACTCTTTT
AcTu	-	974	2.76e-07	TTTAAAGAAA	GGGGGGAGGGGGAAC TACCCCTGTTT	TTAGTGGCTT

Sample ID	Analysis	GenBank Accession	Data Availability
F3	Sanger sequencing of LUR	MF055702	GenBank
F4	Sanger sequencing of whole mitochondrial genome	KP089983	GenBank
F5	Sanger sequencing of LUR	MF055703	GenBank
F7	Sanger sequencing of LUR	MF055704	GenBank
F9	Sanger sequencing of LUR	MF055705	GenBank
F10	Sanger sequencing of LUR	MF055706	GenBank
F11	Sanger sequencing of LUR	MF055707	GenBank
F13	Sanger sequencing of LUR	MF055708	GenBank
F15	Sanger sequencing of LUR	MF055709	GenBank
F16	Sanger sequencing of LUR	MF055710	GenBank
F17	Sanger sequencing of LUR	MF055711	GenBank
F19	Sanger sequencing of LUR	MF055712	GenBank
F20	Sanger sequencing of LUR	MF055713	GenBank
F21	Sanger sequencing of LUR	MF055714	GenBank
fRDI01	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
fRDI02	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
fRDI03	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
fRDI04	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
fRDI05	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
fRDI06	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
mRDI01	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
mRDI02	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
mRDI03	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
mRDI04	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
mRDI05	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank
mRDI06	<i>De novo</i> assembly of mtDNA; SNP	PRJNA170478	https://doi.org/10.6084/m9.figshare.4970762.v3 ; GenBank

Primer	Direction	Sequence 5'-3'	amplified	Lenght(bp)
ATP6-1	F	ATTTGTGCTTCTTTTCAGTGTACTTTT	1517 bp	27
ATP6-1	R	TCATCCCAATTACTAGAATAAAACAGG		27
ATP6-3	F	TAAGTGGTTGGTTCCTTAATGTT	1321 bp	24
ATP6-3	R	AACTTTAATAGGGGAGTTCTTCG		23
ATP6-4	F	TTAAGGAAACGGTGTCTTATTTTTG	1379 bp	25
ATP6-4	R	CTAGGAAAACAGATCACTGACAAGA		25
ATP6-5	F	TTAGTTTGGTTGGGGTAATAATAAGG	1479 bp	26
ATP6-5	R	GTTAGCACAAATACAACCTACCACCAT		26
ATP6-6	F	TATAGATTGAAATCAGTTGGGTTTTTC	1259 bp	27
ATP6-6	R	TCAACTACCCTAGATATCCTTACTGC		27
CYTB-1	F	GATCATATAAACCGTAAGCGAATAATG	1433 bp	27
CYTB-1	R	ATATCTAGCAACTAACCCAACCCTAAT		27
ATP8-2	F	GGCGTGTAGATGAATAGATTTTTA	1255 bp	26
ATP8-2	R	CTCTAAAGAAGGGATCACACCATAA		26
ND6-1	F	TAAATGTAATAAGGCGCTATAATCACC	1242 bp	28
ND6-1	R	TATTTCTGCCACCCTAACAAAATAAAG		27
COX3-2	F	TTAAATTAACCTTAACCAACGAAATAGG	1285 bp	29
COX3-2	R	TAATTGCAATTTGCTTTTTATTAAGTGA		29
ATP8-ND6	F	GCTTTAGGGTGTTCGATCACTCAGGA	1323 bp	27
ATP8-ND6	R	AAACGCCCCCGTAAAAGCTAAGAACAC		27
ATP6-CYTB	F	AGTTGTGTGGATCTGGCCACAGAGAAA	1300 bp	27
ATP6-CYTB	R	TACCTGGGAAACCCCCATTATTCGCTT		27
ND6COX3	F	ATGGCGTAATGGAGGGTGTACGATTC	1218 bp	27
COX3ATP6	R	CACCGGCCATTTAGAAATTCAGGCA		26
CYTB-2	F	TCTGTGTTCAAAAATGTACAGCATAAT	1318 bp	27
CYTB-2	R	CTTTCTTTTTCAGAGACAAGCAACTT		26
ATP6-2R	F	ATTAATAGTAGGTTGGGATGGTTTAGG	1501 bp	27
ND6COX3-R1	R	AAATCGCCCCGGCGTTCTTGAATACTGA		27
ATP62R	F	TTCATGTAATTTTAGGAATTTGCTTTC	2283 bp	27
COX3-R1	R	AAATCGCCCCGGCGTTCTTCTTGAATACTGA		30
ATP6-3R	F	AACTTTAATAGGGGAGTTCTTCG	1729 bp	23
ND6COX3-R1	R	AAATCGCCCCGGCGTTCTTGAATACTGA		28
ATP6-3R	F	AACTTTAATAGGGGAGTTCTTCG	2028 bp	23
COX3-R1	R	TTCATGTAATTTTAGGAATTTGCTTTC		27

ATP6CYTB-F	F	AGTTGTGTGGATCTGGCCACAGAGAAA	1418 bp	28
CYB-R1	R	ATATCTAGCAACTAACCCAACCCTAAT		28
ATP6-5	F	TTAGTTTGGTTGGGGTAATAATAAGG	2076 bp	26
ATP6-6	R	TCAACTACCCTAGATATCCTTTACTGC		28
NewATP8	F	AAGGGGAGGTAGCGAGAAAA	729 bp	20
NewND4	R	GGCAACGAGGAACCTACAGT		20
New16S	F	GTGACACGGTGGATTATTGCTT	1256 bp	22
NewATP6	R	AAACCCCATCACACAAACACAC		22
NewCYTB1	F	TGGGTACATGTCCCGTAGAAGA	870 bp	22
New16S	R	TATGAACGCCTTACCCTATCCC		22
NewND5	F	TGGGGCTCTTCAATAGCTGT	1731 bp	21
NewATP8	R	GCAAGCAAGAGGAGCAAAC		20
NewCR-F3	F	TGTAGAAATAGGCTGAATTCGAGG	1188 bp	24
NewCR-F4	R	ATAACTTTTGC GGCCCTTAGTC		22
NewCR-F4	F	TTGTGATAACTGCTAGGGTGGT	1529 bp	22

Species	GenBank Acc. No.	BRIGS	Easyfig	MEME	MitoPhast
<i>Acanthocardia tuberculata</i>	NC_008452			x	
<i>Arctica islandica</i>	NC_022709			x	x
<i>Coelomactra antiquata</i>	NC_021375			x	
<i>Fulvia mutica</i>	NC_022194			x	
<i>Hiatella arctica</i>	NC_008451			x	
<i>Loripes lacteus</i>	NC_013271			x	
<i>Lucinella divaricata</i>	NC_013275			x	
<i>Lutraria rhynchaena</i>	NC_023384			x	
<i>Meretrix lamarckii</i>	NC_016174	x	x		
<i>Meretrix lamarckii</i> F-type	KP244451			x	x
<i>Meretrix lamarckii</i> M-type	KP244452			x	x
<i>Meretrix lusoria</i>	NC_014809	x		x	x
<i>Meretrix meretrix</i>	NC_013188	x		x	x
<i>Meretrix petechialis</i>	NC_012767	x		x	x
<i>Moerella iridescens</i>	NC_018371			x	
<i>Nuttallia olivacea</i>	NC_018373			x	
<i>Paphia amabilis</i>	NC_016889	x	x	x	x
<i>Paphia euglypta</i>	NC_014579	x		x	x
<i>Paphia textile</i>	NC_016890	x		x	x
<i>Paphia undulata</i>	NC_016891	x		x	x
<i>Ruditapes philippinarum</i> F-type	AB065375	x	x	x	x
<i>Ruditapes philippinarum</i> M-type	AB065374	x		x	x
<i>Semele scabra</i>	NC_018374			x	
<i>Sinonovacula constricta</i>	NC_011075			x	
<i>Solecurtus divaricatus</i>	NC_018376			x	
<i>Solen grandis</i>	NC_016665			x	
<i>Solen strictus</i>	NC_017616			x	
<i>Soletellina diphos</i>	NC_018372			x	
<i>Strongylocentrotus purpuratus</i>	NC_001453			x	

Supplementary Table 4 Most significant GO terms associated with the two DNA motifs found in the LUR. BP = Biological Process; CC = Cellular Component; MF = Molecular Function.

Motif 1	Motif 2
Positive regulation of transcription from RNA polymerase II promoter (BP)	Transcription (BP)
Transcription (BP)	Negative regulation of transcription from RNA polymerase II promoter (BP)
Negative regulation of transcription from RNA Polymerase II promoter (BP)	-
Transcription factor complex (CC)	-
Transcription activator activity (MF)	-

Chapter 3:

Comparative transcriptomics in two bivalve species offers different perspectives on the evolution of sex-biased genes

Comparative transcriptomics in two bivalve species offers different perspectives on the evolution of sex-biased genes.

Journal:	<i>Genome Biology and Evolution</i>
Manuscript ID	GBE-170916
Manuscript Type:	Research Article
Date Submitted by the Author:	12-Sep-2017
Complete List of Authors:	Ghiselli, Fabrizio; University of Bologna, Department of Biological, Geological, and Environmental Sciences Iannello, Mariangela; University of Bologna, Department of Biological, Geological, and Environmental Sciences Puccio, Guglielmo; University of Bologna, Department of Biological, Geological, and Environmental Sciences Chang, Peter; University of Southern California, Biological Sciences – Program in Molecular and Computational Biology Plazzi, Federico; University of Bologna, Department of Biological, Geological, and Environmental Sciences Nuzhdin, Sergey; University of Southern California, Biological Sciences – Program in Molecular and Computational Biology Passamonti, Marco; University of Bologna, Department of Biological, Geological, and Environmental Sciences
Keywords:	RNA-Seq, transcription level, evolutionary rate, gametogenesis, maternal genes, immunity


 SCHOLARONE™
Manuscripts

1
2
3
4
5 Title page
6
7
8

9 **Comparative transcriptomics in two bivalve species offers**
10 **different perspectives on the evolution of sex-biased genes.**
11
12
13
14
15

16 Ghiselli F^{1,†,*}, Iannello M^{1,†}, Puccio G¹, Chang PL², Plazzi F¹, Nuzhdin SV^{2,†},
17 Passamonti M^{1,†}.
18
19
20
21
22

23 Affiliations:

24 ¹Department of Biological, Geological, and Environmental Sciences – University
25 of Bologna, Italy
26
27

28 ²Department of Biological Sciences, Program in Molecular and Computational
29 Biology – University of Southern California, Los Angeles, USA
30
31
32

33
34 [†] These authors contributed equally to this work.
35
36

37
38 *Author for Correspondence: Fabrizio Ghiselli, Department of Biological,
39 Geological, and Environmental Sciences, University of Bologna, Italy.
40 Telephone: +39 051 2094203, email address: fabrizio.ghiselli@unibo.it
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Data Deposition

R. decussatus short reads and transcriptome assembly are available on NCBI (BioProject PRJNA170478).

R. philippinarum short reads and transcriptome assembly are available on NCBI (BioProject PRJNA68513).

The pipeline used for transcriptome annotation (paper in preparation) is available as an Open Science Framework (OSF) project at: https://osf.io/cdkb9/?view_only=f0b2cde926db43719f3d705012c4eeaa

All data (assemblies, differential transcription analysis, annotation, dN/dS, FPKM, clusters of orthologous genes and GO terms) are available on figshare: <https://figshare.com/s/3c0bd3b82e72f882a772>

Abstract

Comparative genomics has become a central tool for evolutionary biology, and a better knowledge of understudied taxa represents the foundation for future work. In this study we characterized the biological processes represented in male and female mature gonads of the European clam *Ruditapes decussatus*, compared to those in the Manila clam *Ruditapes philippinarum* providing, for the first time in bivalves, information about transcription dynamics and sequence evolution of sex-biased genes. In both the species we found a low number of sex-biased genes, probably due to the absence of sexual dimorphism. *R. decussatus* shows a prevalence of female-biased transcripts, the opposite is true for *R. philippinarum*. The transcriptional bias is maintained in only 14% of the orthologs between the two species, and female-biased genes show the highest divergence in transcription. Genes not maintaining the sex bias between the two species are involved in regulatory processes. The dN/dS between orthologs is low, indicating purifying selection, and genes having male-biased transcription in both species evolve significantly faster. Overall, among sex-biased orthologs we observed quite variable transcription opposed to high sequence conservation; regulatory genes show either high transcriptional variability or fast sequence evolution. In contrast with other studies that reported a negative correlation between transcription level and evolutionary rate, our analysis did not find any. We also report the presence of transcripts involved in embryo development in both female and male gametes, and an enrichment of GO terms related to immune response among female-biased genes in *R. philippinarum*.

Key Words

RNA-Seq, transcription level, evolutionary rate, gametogenesis, maternal genes, immunity.

Introduction

Despite the differences in terms of sexual dimorphism and behavior, males and females share almost the same genome, especially in species lacking heteromorphic sex chromosomes. Therefore, the vast majority of sex-specific characters and traits are the result of differential expression of the so-called ‘sex-biased genes’ (Ellegren & Parsch 2007; Parsch & Ellegren 2013; Ranz et al. 2003). The study of sex-biased gene expression is crucial for understanding the mechanisms of gene regulation and evolution (Grath & Parsch 2016): several works investigated the amount of sex-biased genes among animals, showing that the proportion of these genes is extremely variable, depending on the organism, analyzed tissue, developmental and reproductive stage. It has been reported that the number of transcribed sex-biased genes is higher in gonads, since most of them are involved in sexual dimorphism and competition (Parisi et al. 2003; Mank et al. 2010; Harrison et al. 2015). Also, genes that are more or exclusively transcribed in males (male-biased genes) show a higher rate of protein evolution—calculated as the ratio of nonsynonymous to synonymous nucleotide substitution (dN/dS)—as reported in many organisms such as insects, nematodes, birds, and mammals (Grath & Parsch 2012; Meiklejohn et al. 2003; Khaitovich et al. 2005; Pröschel et al. 2006; Zhang et al. 2007; Assis et al. 2012; Harrison et al. 2015; Xu Wang et al. 2015). Even if female-biased genes did not receive the same attention of male-biased genes, some studies conducted in mammals, birds, fish, and insects reported evidence of high dN/dS of these transcripts compared to unbiased genes, namely genes showing no differential expression between sexes (Swanson et al. 2004; Yang et al. 2016; Mank et al. 2007). It is not clear whether the high rate of protein sequence evolution of sex-biased genes, and particularly male-biased genes, is an outcome of positive or relaxed selection. In the literature there are several studies supporting either one or the other theory. Works carried out mostly in *Drosophila* seem to point out that male-biased genes undergo evolution by positive selection (Swanson et al. 2004; Pröschel et al. 2006; Zhang & Parsch 2005): according to this theory, male-male competition drives a faster evolution of male reproductive proteins (Swanson & Vacquier 2002; Clark et al. 2006; Turner & Hoekstra 2008). On the other hand, several studies support the hypothesis that male-biased genes are more dispensable (Mank & Ellegren 2009) and thus under relaxed selection,

1
2
3 while female-biased genes are under stronger constraints due to their functional
4 pleiotropy (Zhang et al. 2007; Duret & Mouchiroud 2000; Xu Wang et al. 2015;
5 Dapper & Wade 2016; Mank et al. 2008). Accordingly, genes exclusively
6 expressed in males showed a higher accumulation of deleterious mutations
7 (Gershoni & Pietrokovski 2014). Finally, the comparison of sex-biased genes
8 across species revealed a large variability in transcription level, suggesting that
9 differences in regulation of sex-biased genes may have a fundamental role in
10 speciation (Romero et al. 2012; Brawand et al. 2011). Particularly, male-biased
11 genes seem to be the most divergent also in terms of transcription level (Torgerson
12 et al. 2002; Zhang et al. 2004, 2007; Meiklejohn et al. 2003; Ranz et al. 2003;
13 Khaitovich et al. 2005): this evidence inspired the hypothesis of a positive
14 correlation between the evolution of protein sequences and transcriptional
15 divergence (Nuzhdin et al. 2004; Lemos et al. 2005; Khaitovich et al. 2005; Sartor
16 et al. 2006; Liao & Zhang 2006). Nevertheless, this pattern is not consistent
17 (Jordan et al. 2004; Harrison et al. 2015; Tirosh & Barkai 2008), indicating that
18 protein sequence evolution and transcription level divergence can be decoupled.
19 Therefore, many questions about evolution of sex-biased genes still remain open.

20
21 What shapes the rate of protein sequence change is a central question for
22 understanding molecular evolution. Several studies have reported different
23 determinants that could influence dN/dS such as, for example, the functional
24 importance of a protein, expression breadth among tissues, pleiotropy, protein-
25 protein interaction, and secondary structure (Larracunte et al. 2008; Ridout et al.
26 2010). Nevertheless, according to the most recent theories, transcription level has
27 been proposed to be the main responsible for the rate of protein evolution (see
28 Zhang & Yang 2015 for a review). Particularly, a strong negative correlation,
29 defined E-R correlation, was found between dN/dS and transcription level, across
30 the three domains of life. One of the main hypothesis to explain the E-R correlation
31 is that highly transcribed genes evolve more slowly thus reducing the amount of
32 misfolded proteins, known to be cytotoxic and damaging for organism fitness
33 (Drummond et al. 2005).

34
35 The development of High-Throughput Sequencing has significantly increased the
36 capability to get insights into the molecular mechanisms driving evolution.
37 Particularly, RNA-Seq produces a large amount of data about both evolution of
38 protein sequence and transcription level, also for nonmodel organisms. The latter
39 point is important, because our knowledge is still restricted to a very limited
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

number of taxa; it is indeed not recommendable to formulate hypotheses and infer general evolutionary patterns based only on a small number of species. For example, although Mollusca is the second phylum of the animal kingdom for number of species, no comparative studies on the relationship between protein evolution and transcription have been performed so far in this group.

In this work we obtained the gonadal transcriptome from males and females of the European clam *Ruditapes decussatus* (Linnaeus, 1758)—also known as grooved carpet shell—and we compared it with the available gonadal transcriptome data (Ghiselli et al. 2012) from the species *Ruditapes philippinarum* (Adams and Reeve, 1850). *R. decussatus* is a bivalve species of the family Veneridae, native to the Mediterranean and European Atlantic coasts. The fishing of *R. decussatus* has historically had a main role in the production of seafood in Italy, Spain and Portugal. The recent introduction in Europe of *R. philippinarum*—native from Philippines, Korea, and Japan—led to a replacement of *R. decussatus* with *R. philippinarum* for aquaculture purposes. Indeed, compared to *R. decussatus*, the Manila clam *R. philippinarum*, reaches sexual maturation at a smaller size, is faster growing, have a greater number of spawning events, a more extended breeding period, and a higher resistance to disease (Ghiselli et al. 2017 and references therein). All these issues probably contributed to a population decline of *R. decussatus* in the Southwestern Europe, as recently reported by Arias-Pérez et al. (2016). Here we characterize the biological processes represented in male and female mature gonad in *R. decussatus*. The comparison with gonad transcription in *R. philippinarum* allows to investigate, for the first time in two bivalve species, the evolution of both protein sequence and transcription level divergence of sex-biased genes. Our analyses provided further insight into the relationship between rate of protein evolution and transcription level.

Materials and Methods

Library preparation

The 12 samples of *R. decussatus* used for this study were collected from the Northern Adriatic Sea, in the river Po delta region (Sacca di Goro, approximate GPS coordinates: 44°50'06"N, 12°17'55"E) at the end of July 2011, during the spawning season. Six males and six females were used to obtain the RNA-Seq

1
2
3
4 library. Clams were sexed by microscope inspection of gonadal liquid collected
5 with a glass capillary tube. The samples were immediately frozen in liquid
6 nitrogen, and preserved at -80°C . Total RNA extraction, mRNA purification and
7 fragmentation, cleaning, and cDNA synthesis were carried out following the
8 protocols of Mortazavi et al. (2008), and modifications as reported in Ghiselli et al.
9 (2012). Samples were barcoded and sequenced over two lanes (for two technical
10 replicates) of an Illumina Genome Analyzer Iix machine, using 76bp paired-end
11 reads.
12
13
14
15
16

17 *De novo* assembly and differential transcription analysis

18 Filtering, *de novo* assembly, and differential transcription analysis were performed
19 following the protocol described in Ghiselli et al. (2012).
20 Assembly completeness assessment was performed using BUSCO v2 (Simão et al.
21 2015) as implemented in gVolante (Nishimura et al. 2017).
22
23
24
25
26

27 Sex-biased transcription

28 For each locus of *R. decussatus* and *R. philippinarum*, we calculated the median
29 FPKM [Fragments Per Kilobase of transcript per Million mapped reads,
30 (Mortazavi et al. 2008)] in males (m_median) and females (f_median).
31 Transcription level fold change between sexes (FC) was obtained as $1 - (f_median$
32 $/ m_median)$. Since many loci presented a sex-biased transcription, we classified
33 genes basing on their m_median, f_median, FC and sex p-value. We defined as
34 “male-specific” those loci with m_median > 1 FPKM, f_median < 1 FPKM, and
35 sex p-value ≤ 0.05 , and as “female-specific” those loci with m_median < 1 FPKM,
36 f_median > 1 FPKM, and sex p-value ≤ 0.05 . We also considered “male-enriched”
37 those loci with FC < -1 and sex p-value ≤ 0.05 ; conversely, loci with FC > 1 and
38 p-value ≤ 0.05 were considered as “female-enriched”. The generic term “male-
39 biased” includes both male-specific and male-enriched genes; similarly, the term
40 “female-biased” includes both female-specific and female-enriched genes. All
41 remaining loci were considered as “unbiased”.
42
43
44
45
46
47
48
49
50
51
52

53 Transcriptome annotation

54 In order to compare the transcriptome of *R. decussatus* with the closely-related
55 species *R. philippinarum*, we used *de novo* assembly and transcription data of *R.*
56
57
58
59
60

1
2
3
4 *philippinarum* from Ghiselli et al. (2012). Both transcriptomes were annotated
5 using a pipeline specifically developed for nonmodel organisms (protocol,
6 information, and data available here:
7 https://osf.io/cdkb9/?view_only=f0b2cde926db43719f3d705012c4eeaa).
8
9

10 11 GO enrichment

12
13 We used the Bioconductor version 3.5 of the package topGO (Alexa et al. 2006) to
14 find GO term enrichment based on Kolmogorov-Smirnov test. GO enrichment was
15 performed for the whole transcriptomes of *R. decussatus* and *R. philippinarum*.
16 Furthermore, we obtained the enriched GO terms of male-specific, male-enriched,
17 female-specific, and female-enriched genes in both species. Data were analyzed
18 with REVIGO (Supek et al. 2011) and GO term networks were visualized using the
19 application DyNet (Goenawan et al. 2016) from the Cytoscape App (Lotia et al.
20 2013).
21
22
23
24
25
26

27 28 Transcription bias of orthologous genes

29
30 Orthologs between the two species were found using OrthoVenn (Yi Wang et al.
31 2015) with the default parameters. In order to investigate whether transcription
32 sex-bias is maintained between orthologous genes, we selected groups of orthologs
33 with at least a sex-biased gene in either species. Since some groups of orthologs
34 included two or more paralogs from the same species with different sex-biased
35 transcription, we defined the SCALE (Shifting in CAtegorical LEvels) index to
36 quantify the overall bias in the transcription. This was computed as follows:
37 paralogs from the same species were categorized into five levels (female-specific,
38 female-enriched, unbiased, male-enriched, and male-specific) and normalized over
39 the total number of paralogs. The cumulative sum of these five frequencies (in the
40 above mentioned order) was then calculated, yielding a number comprised between
41 1 (all paralogs are male-specific) and 5 (all paralogs are female-specific). Finally,
42 this number was divided by 2 and 1.5 was subtracted from the result, in order to
43 obtain a SCALE index (S) comprised between -1 and +1. Formally,
44
45
46
47
48
49
50
51
52

$$53 \quad S = \frac{FS + (FE + FS) + (U + FE + FS) + (ME + U + FE + FS) + 1}{2} - 1.5 =$$

54
55
56
57
58
59
60

$$= \frac{FS + (FE + FS) + (U + FE + FS) + (ME + U + FE + FS)}{2} - 1$$

Where:

FS = frequency of female-specific genes in the group of paralogs;

FE = frequency of female-enriched genes in the group of paralogs;

U = frequency of sex-unbiased genes in the group of paralogs;

ME = frequency of male-enriched genes in the group of paralogs;

MS = frequency of male-specific genes in the group of paralogs.

Using the SCALE index it is possible to quantify biases for groups of orthologs as follows:

$-1 \leq S < -0.6$ male-specific groups of paralogs

$-0.6 \leq S < -0.2$ male-enriched groups of paralogs

$-0.2 \leq S \leq +0.2$ unbiased groups of paralogs

$+0.2 < S \leq +0.6$ female-enriched groups of paralogs

$+0.6 < S \leq +1$ female-specific groups of paralogs

We compared the *S* indexes of groups of paralogs belonging to the same orthologous groups in the two species. Comparisons were visualized in a bubble plot and grouped together in a hierarchical cluster. The GO annotations of most representative clusters were visualized with REViGO.

Rate of protein evolution

To investigate the rate of protein evolution, we selected clusters of orthologs with a single sequence for each species (1:1 orthology). Protein sequences were aligned with MUSCLE (Edgar 2004) and we used the EMBOSS package ‘distmat’ (Rice et al. 2000) to calculate amino acid p-distance. Amino acid alignments were back-translated into nucleotides using a custom R script, and KaKs_Calculator 2.0 (Wang et al. 2010) was used to obtain the ratio of nonsynonymous to synonymous nucleotide substitution (dN/dS). We plotted the dN/dS distribution of orthologs with unbiased transcription in both species (unbiased/unbiased). Additionally, we plotted the dN/dS distribution of orthologs with sex-biased transcription in either

1
2
3 species (unbiased/male-biased; unbiased/female-biased) and orthologs with sex-
4 biased transcription in both species (male-biased/male-biased; female-
5 biased/female-biased; male-biased/female-biased). The Dunn test with the
6 Bonferroni correction was carried out with the R package ‘dunn.test’ to compare
7 the dN/dS distribution among these groups. The same groups were also adopted to
8 analyze the correlation between dN/dS and amino acid p-distance. Finally, in order
9 to investigate the relationship between rate of protein sequence evolution and
10 transcription level, we plotted \log_2 (dN/dS) vs \log_2 (FPKM) for all orthologous
11 genes in both *R. decussatus* and *R. philippinarum*.
12
13
14
15
16
17
18

19 Singlet genes

20
21 We defined as ‘singlets’ those loci that were not recognized by OrthoVenn as
22 orthologs between the two species. We performed a GO term enrichment analysis
23 of singlets, in order to infer the function of such putatively species-specific loci. In
24 addition, we visualized with REViGO the most recurring GO terms in both the
25 species.
26
27
28
29

30 Results

31 *De novo* assembly

32
33
34 More than 67 million paired-end reads were generated from the Illumina
35 sequencing. Both raw reads and transcriptome assembly are available on NCBI
36 (BioProject PRJNA170478).
37
38
39

40
41 The *de novo* assembly yielded 69,279 contigs. The median length of contig
42 sequences is 795 bp and the N50 length is 2,064 bp. Since many of these contigs
43 represent isoforms, they were collapsed in 39,467 representative loci, with median
44 length of 668 bp and N50 of 1,672 bp (see Supplementary Table 1).
45
46

47
48 Supplementary Table 2 shows the results of the completeness assessment
49 performed with BUSCO v2 (briefly: Eukaryota ortholog set: 91.09% complete,
50 96.37% complete + partial, 3.63% missing; Metazoa ortholog set: 87.22%
51 complete, 94.68% complete + partial, 5.32% missing).
52
53
54
55
56
57
58
59
60

Differential transcription analysis

We found 3,935 (10%) loci with a sex-biased transcription in *R. decussatus* and 2,114 (9.2%) in *R. philippinarum*. In *R. decussatus* 661 loci are transcribed more in males than in females (male-enriched), while 1,589 loci are transcribed more in females than in males (female-enriched). We also found 775 loci transcribed only in males (male-specific) and 910 only in females (female-specific). In *R. philippinarum* 909 loci are male-enriched, 631 are female-enriched, 448 are male-specific and 126 are female-specific (Table 1, Figure 1). In total, 60% of sex-biased genes are female-biased in *R. decussatus*, while 60% of sex-biased genes are male-biased in *R. philippinarum*.

Annotation

Of the 39,467 loci in *R. decussatus*, 14,315 (36.3%) were annotated using amino acid similarity (as implemented in our annotation pipeline, see https://osf.io/cdkb9/?view_only=f0b2cde926db43719f3d705012c4eeaa) and 13,697 (34.7%) with nucleotide similarity. GO terms were assigned to 13,865 loci (35.1%). In total, 28,022 loci (71%) were annotated with at least one method, while 11,445 (29%) did not get any annotation. We also re-annotated *R. philippinarum*: of the 22,886 assembled loci, 12,371 (54%) obtained an amino acid-level annotation, 3,997 (17.5%) a nucleotide-level annotation, and GO terms were assigned to 12,064 loci (52.7%). A total of 16,436 loci (71.8%) were annotated with at least one method, while 6,450 loci (28.2%) did not receive any annotation (Table 2).

GO enrichment

We performed a GO enrichment analysis of *R. decussatus* and *R. philippinarum* gonadal transcriptomes. In both the species, the most represented GO terms are involved in biological processes as cell proliferation, meiotic cell cycle, regulation of transcription, regulation of translation, biosynthetic process, chromosome organization, cellular component organization, protein folding, and reproduction (see Supplementary Material 1 and Supplementary Material 2 for the complete results of the GO enrichment analysis of *R. decussatus* and *R. philippinarum*, respectively).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

In order to detect the Biological Processes associated with sex-biased loci, we retrieved the most represented GO terms in male-specific, male-enriched, female-specific, and female-enriched loci. Among male-specific loci of both *R. decussatus* and *R. philippinarum* we found enrichment in GO terms such as sexual reproduction, spermatogenesis, biosynthetic process, and gene expression.

Male-enriched loci are characterized by GO terms involved in reproduction, spermatid development, and nucleic acid metabolism in both species.

Female-specific loci are mainly involved in regulation of metabolic process, transmembrane transport, and lymphocyte and leukocyte differentiation in *R. decussatus*, and in regulation of transcription, and biosynthetic process in *R. philippinarum*.

In female-enriched loci of *R. decussatus*, GO terms are enriched in cytoskeleton organization, regulation of gene expression, regulation of developmental process, cell differentiation, and mitotic cell cycle. In *R. philippinarum*, such loci are principally involved in macromolecule metabolic process, rRNA modification and processing, gene expression, chromatin organization, organelle organization, transcription, DNA replication and recombination, and nucleosome assembly.

Differential transcription and sequence evolution of orthologous genes

A total of 7,180 orthologous groups were found between *R. decussatus* and *R. philippinarum*; 1,521 of these groups (21.2%) include at least a sex-biased gene in one of the two species. The transcription bias of orthologous genes was defined using the SCALE index (see Materials and Methods). We performed a cluster analysis using the SCALE index calculated for each group of orthologs in order to investigate differences and similarities in transcription sex bias between *R. decussatus* and *R. philippinarum* (Figure 2, Table 3). Among the 1,521 groups with at least one sex-biased gene, only in 213 (14%) the sex bias is maintained (Figure 2 cluster B; table 4), while in the remaining 1,308 (86%), the orthologs show a change in sex bias between the two species. More in detail, 521 groups of orthologs (34.2%) are unbiased in *R. philippinarum* and female-enriched in *R. decussatus* (Figure 2, cluster A), 201 groups of orthologs (13.2%) are unbiased in *R. decussatus* and female-enriched in *R. philippinarum* (figure 2, cluster C), 175 groups of orthologs (11.5%) are male-enriched in *R. decussatus* and unbiased in *R. philippinarum* (figure 2, cluster D), and 120 groups of orthologs (7.9%) are

1
2
3 unbiased in *R. decussatus* and male-enriched in *R. philippinarum* (figure 2, cluster
4 E). Supplementary Material 3 reports the GO annotation of the five most abundant
5 clusters (A-E, see Figure 2 and Table 3) of the 1,521 groups of sex-biased
6 orthologs. Among these five clusters, the most recurring GO terms are gene
7 expression, transport, signal transduction, regulation of transcription, translation,
8 DNA replication, RNA processing, regulation of cell cycle, and protein
9 phosphorylation (Supplementary Material 3). We focused on six genes that showed
10 a reversed transcription sex-bias between the two species (Supplementary Table 3):
11 BLAST annotation revealed a pre-mRNA processing factor, innexin, and E3
12 ubiquitin-protein ligase MARCH2, while the remaining three are involved in
13 mitochondrial biology.

14
15 Figure 3 shows the distribution of dN/dS among orthologous genes with 1:1
16 orthology between the two species (N=6,954). We found that the dN/dS
17 distribution of orthologous genes with unbiased transcription in both the species
18 (N=5,508) is not statistically different from almost all groups with a sex-biased
19 transcription in at least one of the two species (Figure 3), indeed the median dN/dS
20 is always included between 0.05 and 0.06. The only condition that shows a
21 different distribution is represented by the orthologs in which male-biased
22 transcription is maintained in the two species (Dunn test p-value=0). This group
23 (N=194) is characterized by a first peak of density corresponding to dN/dS=0.07
24 and a second peak at 0.2 (Figure 3, black line). Genes with a reversed sex-bias
25 between the two species present a median dN/dS of 0.08, but given the low sample
26 size (N=4) we did not include this group in the statistical analysis. The distribution
27 of amino acid p-distance (Supplementary Figure 1) shows that orthologous genes
28 between *R. decussatus* and *R. philippinarum* are characterized by a median
29 divergence of 8%. In only 1.4% of genes the p-distance is $\geq 60\%$. We investigated
30 the relationship between dN/dS of orthologs detected by OrthoVenn and the amino
31 acid p-distance in both unbiased genes and sex-biased genes. We found that among
32 orthologs with lower amino acid divergence, approximately below 40%, the
33 relationship between dN/dS and p-distance shows a linear trend (Figure 4a, black
34 dashed line). Instead, when the amino acid divergence is higher, the trend is better
35 described by an exponential function (Figure 4a, red dashed line). This pattern is
36 particularly evident in unbiased genes (Figure 4b). On the contrary, we found that
37 among sex-biased categories, genes with a p-distance higher than 40% are rare, a
38 linear model fits the data (Figure 4c-f, colored solid lines), and the trend for each
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 sex-biased category is comparable to the linear trend of unbiased genes (Figure 4b-
5 f, black dashed lines). Among the genes following an exponential relationship
6 between dN/dS and amino acid p-distance (Figure 4a, red dashed line), we focused
7 on two clusters: one includes orthologs with a p-distance included between 40-60%
8 and a dN/dS<0.2, the other includes orthologs with a p-distance >60% and a
9 dN/dS>0.2. We refer to the orthologs belonging to the abovementioned clusters as
10 'fast-mutating' and 'fast-evolving', respectively (see Discussion). The most
11 recurring GO terms associated with the two clusters are reported in Supplementary
12 Material 4. Table 4 shows the top 20 most frequent GO terms associated with 'fast-
13 mutating' and 'fast-evolving' orthologs. Of the 25 GO terms represented, 15
14 appear in both groups of orthologs (metabolic process, transport, oxidation-
15 reduction process, transcription, DNA-templated, regulation of transcription DNA-
16 templated, phosphorylation, cellular protein modification process, transmembrane
17 transport, signal transduction, translation, ion transport, carbohydrate metabolic
18 process, proteolysis, protein phosphorylation, intracellular signal transduction),
19 while 10 appear only in one group (biosynthetic process, response to stimulus,
20 regulation of RNA biosynthetic process, regulation of nucleic acid-templated
21 transcription, and ion transmembrane transport for the 'fast mutating' orthologs;
22 nucleobase-containing compound metabolic process, lipid metabolic process,
23 nucleotide biosynthetic process, cellular response to DNA damage stimulus, and
24 nucleic acid phosphodiester bond hydrolysis for the 'fast-evolving' orthologs).
25
26 Finally, by analyzing the relationship between dN/dS of orthologous genes and
27 FPKM in *R. decussatus* (Figure 5), we found no evidence of correlation between
28 rate of protein evolution and transcription level (Spearman's rank correlation $\rho=-$
29 0.01; p-value=0.3). The same lack of correlation was detected in *R. philippinarum*
30 (Spearman's rank correlation $\rho=-$ 0.03; p-value=0.001).
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

46 GO annotation of singlets

47
48 For 5,652 loci of *R. decussatus*, OrthoVenn did not find any ortholog in *R.*
49 *philippinarum*, while 5,066 loci of *R. philippinarum* had no orthologs in *R.*
50 *decussatus*. We performed a GO term enrichment analysis of singlets: the network
51 in Figure 6 shows the biological processes enriched in singlets of *R. decussatus*
52 (green) and *R. philippinarum* (red); GO terms shared between the two species are in
53 white. The complete results of this analysis are shown in Supplementary Material
54
55
56
57
58
59
60

1
2
3
4 5. In *R. philippinarum* singlets seem to be involved mostly in cell part
5 morphogenesis, cAMP biosynthesis, inorganic anion transmembrane transport,
6 lymphocyte activation, leukocyte differentiation, and ribonucleoprotein complex
7 subunit organization. In *R. decussatus* singlets, the most represented GO terms are
8 related to anatomical structure formation involved in morphogenesis, gamete
9 generation, locomotion, protein dephosphorylation, cell development, biological
10 regulation, and reproduction. The only biological process found to be enriched in
11 singlets of both species is cell projection organization.
12
13
14
15
16

17 18 **Discussion**

19
20 In this work we obtained the transcriptome of mature gonads in male and females
21 of *R. decussatus*, and performed a comparative analysis with the related species *R.*
22 *philippinarum*. Since gonads are known to be the tissue with the higher
23 transcription of sex-biased genes (Zhang et al. 2004, 2007; Reinke et al. 2004;
24 Torgerson et al. 2002; Good & Nachman 2005) this experiment gave the
25 opportunity to investigate the evolution of sex-biased genes in two bivalve species,
26 both in terms of protein sequence and transcription level. Also, we report here
27 information about the biological processes represented in male and female mature
28 gonad, as well as an analysis of the relationship between transcription level and
29 rate of protein evolution.
30
31
32
33
34
35

36 About 100 species of bivalve molluscs show the Doubly Uniparental Inheritance
37 (DUI; (reviewed in Zouros 2013), an unusual mechanism of mitochondrial
38 transmission. While *R. philippinarum* has DUI, so far *R. decussatus* did not show
39 evidence for the presence of two sex-linked mitochondrial genomes typical of DUI
40 (Ghiselli et al. 2017), so it is probable that it is a species with strictly maternal
41 inheritance. It is therefore conceivable that some of the diversity in transcription of
42 genes related to mitochondrial biology is related to a different mitochondrial
43 inheritance system between the two species. Given the complexity of the issue, we
44 will deal with such issue in a dedicated manuscript (Iannello et al., in preparation).
45
46
47
48
49
50

51 52 Biological processes represented in mature gonads

53
54 The GO term analysis of the whole transcriptomes shows, for both the species, an
55 enrichment of biological processes involved in reproduction, cell proliferation,
56 meiotic cell cycle, regulation of transcription and translation, chromosome
57
58
59
60

1
2
3
4 organization and cellular component organization, as expected given the analyzed
5 tissue and reproductive stage (Supplementary Material 1, Supplementary Material
6 2). Also the most recurring GO terms associated with sex-biased genes reflect
7 biological processes which are typically involved in cell proliferation, which
8 characterizes both female and male mature gonads. So, it is unsurprising to find
9 among female-biased transcripts an enrichment of GO terms involved in cell cycle
10 process, cell division, and cell project organization. GO terms reflect also the
11 dynamics of cytoskeleton organization: cell divisions are indeed strictly controlled
12 in both time and space from the action of microtubules and microfilaments, that
13 ensure the formation of the spindle and the segregation of homologous
14 chromosomes during the first meiotic division, the segregation of sister chromatids
15 during the second meiotic division, the asymmetry between oocyte and polar
16 bodies, and cell division (Brunet & Maro 2005). Among female-biased transcripts
17 there is also an overrepresentation of GO terms involved in regulation of
18 macromolecule metabolism, gene expression, regulation of transcription and
19 organic substance transport: this is likely due to cytoplasmic maturation of the
20 oocyte, which consists in the accumulation of mRNA, proteins, and nutrients
21 required for early embryo development (Watson 2007). Finally, we found an
22 enrichment of GO terms involved in immune system and embryo development
23 which will be discussed in detail in two dedicated sections. For what concerns
24 male-biased transcripts, they are enriched in GO terms involved in reproduction
25 and cell division as well, matching the biological processes expected for a mature
26 gonad. GO terms associated with organization of cytoskeleton are largely
27 represented: besides mitosis and meiosis, microtubules are heavily involved in
28 spermiogenesis, the last phase of spermatogenesis, where they are necessary for
29 nuclear elongation, and for the development of acrosome and flagellum (Sperry
30 2012). Among male-biased genes we found GO terms related to embryo
31 development, as well (see discussion below).

32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49 Comparative analysis of transcription level and rate of protein evolution of sex-
50 biased orthologous genes between *R. decussatus* and *R. philippinarum*

51
52
53 In both *R. decussatus* and *R. philippinarum*, about 10% of the assembled contigs
54 show a sex-biased transcription (Table 1). These values are lower compared to
55 what reported in other taxa, even when detection of sex-biased genes was
56
57
58
59
60

1
2
3
4 performed in somatic tissues: more than 81% of genes showed a sex-biased
5 transcription in frogs (Malone et al. 2006), about 75% in wasps (Xu Wang et al.
6 2015), up to 57% in *Drosophila* (Ranz et al. 2003), 50% in *Daphnia pulex* (Eads et
7 al. 2007), and similar patterns were found in copepods (Poley et al. 2016),
8 Anopheles (Papa et al. 2017), fish (Small et al. 2009), birds (Mank et al. 2010), and
9 mammals (Yang et al. 2006). Nevertheless, it should be considered that the amount
10 of sex-biased genes increases with the number of tissues analyzed (Ellegren &
11 Parsch 2007; Yang et al. 2006), while only gonads were investigated in this work.
12 Furthermore, clams lack sexual dimorphism as well as mating behavior, which are
13 responsible for the majority of differential transcription between sexes (Ellegren &
14 Parsch 2007; Harrison et al. 2015). Therefore, in these bivalves all the genes
15 showing differential transcription between sexes are likely involved in
16 gametogenesis, and this experiment offers the opportunity to observe the
17 transcriptional difference between female and male gonads and gametes. Since
18 gonads were sampled during the same stage of gametogenesis in two related
19 species that lack sexual dimorphism, we did not expect to detect considerable
20 differences in transcription of sex-biased genes. Nevertheless, while we found that
21 among sex-biased genes of *R. decussatus* there is a higher proportion of female-
22 biased genes—as already seen in *Daphnia galeata* and *Ischnura elegans*
23 (Huylmans et al. 2016; Chauhan et al. 2016)—*R. philippinarum* is characterized by
24 a higher proportion of male-biased genes—which seems to be the most common
25 situation, as reported by the studies cited in the Introduction. Overall, in 86% of the
26 orthologous genes the sex bias is not maintained between the two species (Figure
27 2, Table 3). The most frequent condition is represented by genes that are female-
28 biased in one species and unbiased in the other species; therefore, despite male-
29 biased genes showing the greater transcription divergence in many studies (Ranz et
30 al. 2003; Meiklejohn et al. 2003; Brawand et al. 2011; Khaitovich et al. 2005),
31 female-biased genes are the most transcriptionally variable in these two bivalves, a
32 situation previously reported in frogs (Malone et al. 2006). Among genes where
33 the sex bias is not maintained, the most represented GO terms include gene
34 expression, transport, signal transduction, regulation of transcription, translation,
35 DNA replication, RNA processing, regulation of cell cycle, and protein
36 phosphorylation (Supplementary Material 3). This finding is consistent with genes
37 having divergent transcription being involved in regulatory functions.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Sex-biased genes are considered to evolve faster, in particular those expressed preferentially or exclusively in reproductive tissues (Parisi et al. 2003; Harrison et al. 2015; Mank et al. 2008). In order to test this condition, we investigated the rate of protein sequence evolution in both unbiased and sex-biased genes, and among the latter we separated those where the bias was maintained from those where the bias changed between the two species (Figure 3). We found that the only group of orthologs with a significant higher rate of protein evolution is composed of genes that are male-biased in both the species (Figure 3, black line, Dunn test p -value=0). Conversely, orthologs that show a male-biased transcription in only one of the two species have dN/dS comparable to that of unbiased genes, as already reported in birds and *Drosophila* (Harrison et al. 2015; Grath & Parsch 2012; Metta et al. 2006). Similarly, the dN/dS distribution in female-biased genes is comparable to that in unbiased genes, independently on whether the female-bias was maintained or not in the species. So, if on the one hand male-biased genes show a higher dN/dS, on the other hand female-biased genes are the most variable in terms of transcription between the two species. This evidence reveals two different kinds of evolution of sex-biased genes, where a more rapid evolution of protein sequence seems to be predominant for male-biased genes, whereas a more variable transcriptional regulation is presumable for female-biased genes. Also, this observation is consistent with the assumption that transcription level divergence and rate of protein evolution are decoupled (Jordan et al. 2004; Harrison et al. 2015; Tirosh & Barkai 2008). Why would male- and female-biased genes evolve differently? One possible hypothesis is that female-biased genes are more constrained in terms of protein sequences, because they are involved in more biological processes ('functional pleiotropy', see: Zhang et al. 2007; Mank & Ellegren 2009), but, perhaps for the same reason, are subject to a more variable transcriptional regulation. Following this rationale, male-biased genes would be more specialized, but it is not clear whether their higher amino acid sequence evolution is a result of positive selection, as expected for male proteins involved in sexual competition (Swanson & Vacquier 2002; Clark et al. 2006; Turner et al. 2008), or if they are more dispensable and thus free to accumulate mutations (Gershoni & Pietrokovski 2014). Further analysis, involving more tissues and several different species may help to investigate this point in *Bivalvia*. Nevertheless, in our data, dN/dS distribution is strongly skewed toward 0 in all cases, values higher than 0.4 are quite rare, and we did not find any sign of positive

1
2
3 selection. Given what reported in the literature, this is surprising because proteins
4 involved in male reproductive traits, such as sperm-eggs recognition, or sperm
5 competition are expected to have dN/dS higher than 1, or at least close to this value
6 (Swanson & Vacquier 2002; Dapper & Wade 2016; Clark et al. 2006). This should
7 be even more evident in marine invertebrates where fertilization occurs externally,
8 and positive selection acting on proteins involved in fertilization is thought to
9 trigger the evolution of reproductive isolation (Turner & Hoekstra 2008). We
10 wondered if these results were due to a biological condition or to a technical
11 artefact. On the one hand, if we obtained an accurate representation of the actual
12 evolutionary rates, then orthologs transcribed in the gonads of these two bivalve
13 species experience only purifying selection. So why such a slow evolutionary rate?
14 In species with heteromorphic—thus non-recombining—sex chromosomes, sex-
15 biased genes are non-randomly distributed across the genome, so there can be
16 cooperation and conflict among different chromosomal regions depending on
17 whether they are co-transmitted as part of male or female sex determination. Such
18 transmission asymmetry entails a high probability of sex chromosomes being
19 involved in genomic conflicts, leading to sexually antagonistic variation (Rice &
20 Chippindale 2001; Rice 2013), resulting in faster evolution. In particular, the Y
21 chromosome does not recombine and is male-limited, and males generally evolve
22 faster, so the conflicts will result in a faster evolution of male-biased genes. One
23 could argue that the Y chromosome contains only few genes, but—to use Rice’s
24 words—Y is a “coding dwarf” but a “regulatory giant” (Rice 2013): in *Drosophila*,
25 polymorphisms at loci on Y chromosome influence thousands of genes (Stewart et
26 al. 2010). So given that bivalves lack heteromorphic sex chromosomes, such
27 conflicts—thus evolutionary rates—should be different, perhaps softer. Another
28 non-mutually exclusive hypothesis should be considered. The sex determination
29 system in bivalves is still unknown, but there is experimental evidence that sex is
30 determined by maternal nuclear genome (Zouros 2013), and—since triploids
31 develop male gonads—that maleness is achieved by exceeding a threshold of some
32 yet unknown masculinizing factor. It was proposed that the activation of sex
33 determination genes depends on genetic elements (RNAs, proteins) stored in the
34 oocyte, whose concentration depends on maternal genotype. In this way, F1 sex
35 depends only on maternal genotype, while paternal genotype contributes to F2 sex
36 (see Fig. 7 in Ghiselli et al. 2012 for a scheme). Brisson and Nuzhdin (2008)
37 showed that in a system with a strong reproductive skew between males and
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

females (pea aphids), male-biased genes undergo mutational decay, while female-biased genes are under constant selection. Clams do not experience such a strong reproductive skew, but if the above-mentioned model is true, feminizing factors—which should be female-biased genes—and masculinizing factors—which should be male-biased genes—might experience different selective pressures, mainly because masculinizing genes do not produce their effect in the embryos receiving them, but in their progeny. Basically their transmission does not depend on their masculinization action, and this might result in the elimination of pressure on male-biased genes to evolve fast. This idea is still at the stage of speculation, but we think it deserves further investigation. On the other hand, to explain the unexpectedly low dN/dS in sex-biased genes, we also want to consider the possibility of a too strict method for orthology detection: this could lead to a considerable number of false negatives, especially among the more variable orthologous genes, and the consequent exclusion of such sequences from the analyses. This would be particularly damaging, because fast evolving genes are the more informative and interesting genes for studying molecular evolution. In order to understand whether the method we used to identify orthologs introduced a bias toward slow evolving sequences, we investigated the relationship between dN/dS and amino acid p-distance (Figure 4). The analysis showed that the method inferred orthology in sequences that diverged up to 80%. Nevertheless, orthologs with a p-distance $\geq 60\%$ are about 1% of the total, and they are almost absent among sex-biased genes, that are expected to be the most variable. Given all the published literature on this subject, the complete absence of orthologous genes showing a dN/dS ≥ 0.8 is quite surprising, especially considering that we are analysing genes transcribed in gonads. A possible explanation is that most of the rapidly evolving genes were not recognized as orthologs and were erroneously included among singlets. Detection of orthologs is a well-known problem: to date, basically two approaches are used: the graph-based methods and the tree-based methods (Kristensen et al. 2011). In the graph-based method, clusters of orthologs are based on sequence similarities, but, as mentioned before, the most rapidly evolving and thus informative sequences could be discarded, due to high sequence divergence. On the other hand, a tree-based method requires good *a priori* knowledge of both gene family trees and species trees, that is not easily obtainable especially for nonmodel taxa, which are poorly represented in the construction of gene family trees. Besides, by comparing different programs based on both methods on a

1
2
3 curated database of orthologs, Trachana et al. (2011) found that all repositories
4 predict only a fraction of these orthologs.
5

6
7 Going back to our data, it is worth noting that the relationship between dN/dS and
8 amino acid p-distance has two different trends, depending on whether the p-
9 distance is lower or higher than 40% (Figure 4a). Below 40% the relationship
10 follow a linear trend; on the contrary, when the amino acid divergence is higher
11 than 40%, the correlation is better described by an exponential function. We can
12 think the x-axis on the plot (amino acid p-distance) as a measure of
13 nonsynonymous change, while the y-axis represent the amount of nonsynonymous
14 change “normalized” by the proportion of synonymous change. Therefore, for a
15 fixed p-distance, an increasing dN/dS value corresponds to a decreasing proportion
16 of synonymous changes, namely a faster rate of amino acid sequence evolution.
17 Accordingly, we can observe that between 40% and 60% of p-distance there is a
18 cluster of genes that shows a low dN/dS, meaning that the high number of
19 nonsynonymous changes is coupled with an even higher proportion of synonymous
20 changes. These genes therefore seem to experience a high mutation rate, but a
21 relatively low rate of evolution, intended as the proportion of nonsynonymous
22 change (‘fast-mutating’ orthologs). Between a p-distance of 60% and 80% the
23 dN/dS rate is much higher, meaning that for the genes included in this interval
24 most of the change is nonsynonymous, so they undergo a faster evolution (‘fast-
25 evolving’ orthologs). Unexpectedly, most of such genes did not show a sex-biased
26 transcription in either *R. decussatus* or *R. philippinarum*. The top 20 most recurring
27 GO terms associated with ‘fast-mutating’ and ‘fast-evolving’ orthologs (Table 4)
28 are similar: the two groups share 15 of the 25 GO terms listed (Table 4), and the
29 remaining 10 are closely related among each other. So there seems to be little
30 difference in overrepresented biological processes between the two groups, and
31 orthologs involved in gene expression regulation (at any level) are among the most
32 rapidly evolving genes in our dataset.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

49 The controversial relationship between evolutionary rate and transcription level

50
51 According to recent studies, the evolutionary rate of a protein would be mainly
52 influenced by its transcription level (Zhang & Yang 2015 and references therein),
53 and a negative correlation between transcription level and evolutionary rate,
54 defined E-R correlation, was found across a variety of species. In contrast, our
55
56
57
58
59
60

1
2
3
4 analysis on two bivalves shows no correlation. One hypothesis is that bivalves
5 undergo different patterns of protein evolution and/or that protein sequence
6 evolution is not driven by transcription level, for yet unknown reasons. In this case,
7 bivalves—or at least the two clam species we analyzed here—would represent an
8 exception, and further investigation would be necessary to understand how, and
9 why. A second hypothesis concerns the methodology: is our opinion that the
10 common practice used for investigating the relationship between dN/dS and
11 transcription level could yield inaccurate results, for three main reasons.
12
13

14
15 1) Transcription level is extremely variable. Transcription is a quite noisy
16 process—especially in multicellular eukaryotes—and it is influenced by both
17 genetic factors—that modify gene expression depending, for example, on the
18 tissue, developmental stage, phase of life cycle, etc.—and environmental factors
19 (e.g.: diet, stress). Most of the times, the variation caused by genetics,
20 environment, and by the combination of both is unpredictable. There could also be
21 technical issues, like different sequencing methods and RNA quality, which are
22 known to influence the measurement of transcription level. In addition, in
23 multicellular organism the transcription level is often calculated averaging the
24 mRNA concentration across several tissues: such values are likely inaccurate and
25 not representative of the physiological condition of the organism. In order to
26 perform reliable comparative analyses, it is important to use homogeneous data,
27 and for all the abovementioned reasons, this is a condition which is not often
28 achieved.
29

30
31 2) dN/dS are calculated between the species for which transcription level is
32 measured and one related species. Since the rate of protein evolution is influenced
33 by the phylogenetic distance between the considered species, dN/dS are variable as
34 well, depending on the species used in the comparison. Since available
35 transcriptome data are not equally representative of all the taxa, it is often difficult
36 to find species with homogeneous evolutionary distances, so that the comparisons
37 would be consistent across all the analyzed taxa.
38

39
40 3) For each dN/dS, we do not know whether it is the result of an even
41 accumulation of divergence along the two branches that separate the two species
42 under analysis, or if it is due to just one species evolving much faster than the
43 other. In the latter case, the evolutionary rate is overestimated in one species, and
44 underestimated in the other. In addition, if transcription levels are not strongly
45 correlated between the two species the relationship between dN/dS and
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 transcription level would show different trends, depending on what species is used
4 for the transcription quantification. In such case, what is the correct trend?

5
6 In the present work we tried our best to overcome the problems mentioned in point
7 1 by comparing data obtained using the same experimental design (number of
8 biological and technical replicates, sex, tissue, sampling season, and reproductive
9 phase), the same protocols performed in the same lab, the same sequencing
10 technology, and the same pipeline of analysis. The fact that we used two species
11 and a single tissue eliminated the other issues raised in point 1, and those in point
12 2. Point 3 is more difficult to address: despite *R. decussatus* and *R. philippinarum*
13 are morphologically and ecologically very similar, during their evolutionary
14 history they experienced very different population dynamics (Arias-Pérez et al.
15 2016; Cordero et al. 2017), which could have resulted in significantly different
16 evolutionary rates between the two species. Unfortunately, with the available data,
17 a more accurate estimate is not possible. For what concerns the transcription level
18 of orthologous genes, the two species show a moderate correlation (Spearman's
19 rank correlation $\rho=0.2$, $p\text{-value}=2.2\text{E-}16$, see Supplementary Figure 2). This affects
20 the E-R plots that yield different results depending on which transcription data—*R.*
21 *decussatus* or *R. philippinarum*—are being used in the analysis (Figure 5). In this
22 case, despite the differences, the final result is unambiguous (i.e. no correlation),
23 but in cases where the transcription level of orthologs is less/not correlated, results
24 might be significantly different, potentially even opposite. Thus, the chance of
25 getting ambiguous results increases with the transcriptional divergence between the
26 analyzed species, which, in turn, increases with phylogenetic distance plus a large
27 number of factors which—as discussed above—are difficult to predict or
28 standardise.

29
30 More work is needed to establish whether the absence of E-R correlation reported
31 here is an exception—and what biological/evolutionary causes are responsible for
32 it—or it is a more widespread feature. In the latter case, the hypothesis that
33 transcription level is the main responsible for the rate of protein evolution should
34 be revised. The results of this work are not compelling enough to reject the E-R
35 correlation theory, but in our opinion they show that caution is needed when
36 performing comparative analyses, especially when doing it across a wide range of
37 distantly related species. Consequently, we think that more clear-cut evidence is
38 needed to support the hypothesis by which transcription level drives protein
39 evolution. Alternatively, many other features are thought to be involved in protein
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 evolution, such as functional importance, pleiotropy, protein-protein interaction,
4 network property, and structural constraints (Larracuente et al. 2008; Ridout et al.
5 2010).
6
7

8
9
10 Transcripts involved in embryo development are stored in gametes

11
12 Animal eggs contain the substances required for sustaining the first stages of
13 development. Other than vitamins, minerals, fatty acids, and other nutrients
14 provided by the yolk, egg cells contain RNAs and proteins produced by the so-
15 called ‘maternal genes’ during oogenesis. Such products sustain and guide embryo
16 development, especially before the activation of zygotic gene expression (Marlow
17 2011). Indeed, the first cell divisions after fertilization occur in absence of new
18 transcription, and in this phase the embryonic development relies solely on
19 maternal gene products. The analyses here reported further support the existence of
20 maternal genes in both the species. Indeed, among female-biased genes, we found
21 an enrichment of GO terms involved in embryonic organ morphogenesis, and in
22 development and formation of primary germ layer. This result supports the storage
23 in the egg of maternal gene products (in this case mRNAs) that would be used
24 during embryo development. Interestingly, we also found an enrichment of GO
25 terms involved in embryo development and organ morphogenesis among singlets
26 of both *R. decussatus* and *R. philippinarum*, indicating that a large number of
27 genes involved in such processes are not conserved between the two species.
28

29
30 What about paternal contribution? Until few years ago it was thought that the
31 primary function of sperm was to deliver the paternal DNA to the embryo.
32 Recently, it was discovered in mammals, insects, and plants that sperm carry
33 thousands of RNAs (Hosken & Hodgson 2014; Dadoune 2009). These transcripts
34 persist in spermatozoa, where the machinery for their translation is inactive (Miller
35 et al. 2005), so it is unlikely that the RNAs stored in spermatozoa are necessary for
36 its survival, but it is more plausible that they contribute to embryo development,
37 even if their function is still unknown (Hosken & Hodgson 2014). Accordingly, in
38 the last years, evidence of sperm transcriptional contribution to the offspring
39 development is increasing. Several experiments highlighted the importance of
40 epigenetic inheritance acquired through sperm RNAs (Chen et al. 2016), and the
41 role of paternal miRNAs in the first embryo division (Yuan et al. 2015; Liu et al.
42 2012). Interestingly, among male-biased genes, we found several GO terms
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 involved in developmental process and determination of bilateral symmetry. To the
4 best of our knowledge, this is the first evidence of sperm transcripts related to such
5 processes, supporting the hypothesis that sperm can be involved not only in the
6 first cleavage division, but also it could have a role in the following stages. A role
7 of paternal factors (RNAs, proteins) in embryo development and sex determination
8 has been already proposed for *R. philippinarum* (Ghiselli et al. 2012; Milani et al.
9 2013).

16 Immune system

17
18 Knowledge about invertebrate immunity is rapidly increasing (Yuan et al. 2014):
19 while vertebrates are characterized by an adaptive immune system, where
20 immunological memory allows to develop specific responses to pathogens after
21 their first attack, invertebrates possess only an innate immune system that was
22 initially thought to be a quick, nonspecific defense to all pathogens and
23 distinguished by a lack of memory. More recently, several experiments
24 demonstrated instead that invertebrates show a great plasticity of immune response
25 to both different pathogens and different strains (Kurtz & Franz 2003).
26 Furthermore, there is evidence of immune memory in several invertebrate taxa,
27 that protects organisms from specific pathogens after their secondary exposure
28 (Milutinović & Kurtz 2016; Kurtz & Franz 2003). Although most of the molecular
29 components involved in innate immune system are still unknown, it has been
30 proposed that high genomic diversity, alternative splicing, rearrangement of gene
31 exons, as well as synergistic interaction of components and dosage effects could be
32 responsible for these highly specific responses in invertebrates (Schulenburg et al.
33 2007). Also, it is known that in both vertebrates and invertebrates, maternal
34 immunity plays a crucial role for the survival of early stage embryos (Grindstaff et
35 al. 2003; Knorr et al. 2015). By this mechanism, mothers transfer immunity to the
36 offspring through eggs, so that embryos are protected from external pathogens
37 during the first stages of their development. This is particularly important in
38 species where the fertilization is external and both eggs and embryos are highly
39 exposed to environmental factors. Immune response has been thoroughly
40 investigated in Bivalvia due to the ecological and commercial importance of this
41 group (see for example: Moreira et al. 2012; Pauletto et al. 2014; Gerdol & Venier
42 2015), and evidence of both memory and maternal transfer of immunity were
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

found (Cong et al. 2008; Lingling Wang et al. 2015). Immune system components are recurring in the gonadal transcriptome of both *R. decussatus* and *R. philippinarum* females, and we found several GO terms involved in leukocyte differentiation among female-biased transcripts. Moreover, GO terms involved in immune system were frequent among genes with p-distance between 40-60% and $dN/dS < 0.2$ ('fast-mutating' orthologs), suggesting a high mutation rate but a contained rate of evolution of these genes. Intriguingly, GO terms related to immune response resulted to be enriched among singlets of *R. philippinarum*, but not of *R. decussatus*. We propose two possible explanations for this result: i) these genes are present in both *R. decussatus* and *R. philippinarum* genomes, but they are transcribed only in *R. philippinarum* because of the exposition to a pathogen that was absent in the environment from where *R. decussatus* was sampled. Since immunity is highly specific, the upregulated transcription of genes involved in strain-specific response in *R. philippinarum* females would not be transcribed in *R. decussatus*. Accordingly, it has been reported that the presence of a pathogen can influence the maternal transfer of immunity, so that the offspring receive a strain specific defense from the pathogen present in that environment (Yue et al. 2013; Little et al. 2003). ii) These genes are only present in the genome of *R. philippinarum*, that evolved a higher genetic diversity in response to pathogens, highlighting a very different evolution of immune response between the two species. In a transcriptome comparative study of genes involved in immune system between *R. decussatus* and *R. philippinarum* Moreira et al. (2012) reported that, following bacterial infection, *R. decussatus* seems to have a less effective and lower immune response compared to *R. philippinarum*. This might explain the population decline of *R. decussatus* and its lower resistance to diseases, explaining also the ongoing replacement of the European clam by the invasive *R. philippinarum*, which might have been evolved a more efficient response to pathogens.

Finally, we found an overrepresentation of GO terms involved in lymphocyte activation in both *R. decussatus* and *R. philippinarum*: this is surprising, since lymphocytes are involved in vertebrate immunity. Nevertheless, lymphocyte-like cells were found in amphioxus (Huang et al. 2007) and this could open new perspectives about invertebrate immunity.

Conclusions

This work provides new information about transcription dynamics and sequence evolution of sex-biased genes in a group of Metazoa for which such data are missing. Compared with other taxa, both the bivalve species analyzed showed a low number of sex-biased genes, probably due to the absence of sexual dimorphism and other sex-specific features (e.g. mating behavior). Surprisingly, we found striking differences in transcription between the two species: *R. decussatus* shows a prevalence of female-biased transcripts, while in *R. philippinarum* the majority of sex-biased genes are male-biased. Moreover, the transcriptional bias is maintained in only 14% of the orthologs between the two species, and—contrarily to what reported in multiple studies on other animals—female-biased genes show the highest divergence in transcription. Genes not maintaining the sex bias between the two species appear to be mainly involved in regulatory processes. For what concerns sequence evolution, orthologs showed a low dN/dS indicating a prevalence of purifying selection; genes having a male-biased transcription in both species resulted to be evolving significantly faster than other groups of genes. Among orthologs, we identified two groups that stood out against the others: a cluster of ‘fast-mutating’ genes, and a cluster of ‘fast-evolving’ genes. The biological processes associated to both the groups are involved in regulation of gene expression. Overall, the central theme seems to be that of a quite variable transcription opposed to a high sequence conservation; genes involved in regulatory functions show either high transcriptional variability or fast sequence evolution. We also report the presence of transcripts involved in embryo development in both female and male gametes, and an enrichment of GO terms related to immune response among female-biased genes in *R. philippinarum*. During the development of this work we had to face several technical challenges typical of comparative analyses performed on nonmodel organisms. This allowed us to think about the difficulties in inferring orthology and about some downsides of the common practices used to investigate the relationship between protein sequence evolution and transcription level. The concerns we raised about such technical approaches do not have straightforward solutions, but we think a significant improvement can be achieved through more careful experimental designs—from both methodological and biological points of view—and a wider sampling across the whole ‘Forest of Life’. Our growing understanding of the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

diversity of life clearly advise against the routine of formulating hypotheses and inferring general evolutionary patterns based only on a small number of species. Therefore, since comparative genomics has a fundamental role in almost every field of biology, an improvement in comparative methods will represent one of the main challenges for the next future. Such improvement require a better knowledge of genomes and transcriptomes that, in turn, depends on our ability in annotating genes and inferring phylogenetic relationships across taxa. The problem is evidently circular, and at the moment the focus should be on getting a more uniform representation of the actual biodiversity in genomics data. This is a demanding endeavour, but in the last few years numerous international collaborative projects have been established with the goal of filling the gap of knowledge sequencing an increasing number of species (Voolstra et al. 2017 and references therein).

Acknowledgements

We thank Liliana Milani for providing useful comments and suggestions about the manuscript, and the Waitt Foundation.

Funding

This work was supported by the Italian Ministry of Education, University, and Research (MIUR) FIR Programme no. RBFR13T97A funded to FG, the Canziani Bequest funded to MP, and the NIH grant no. RO1GM098741 funded to SVN.

References

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 22:1600–1607.

Arias-Pérez A et al. 2016. Assessing the geographic scale of genetic population management with microsatellites and introns in the clam *Ruditapes decussatus*. *Ecol Evol*. 6:3380–3404.

Assis R, Zhou Q, Bachtrog D. 2012. Sex-biased transcriptome evolution in *Drosophila*. *Genome Biol Evol*. 4:1189–1200.

- 1
2
3 Brawand D et al. 2011. The evolution of gene expression levels in mammalian
4 organs. *Nature*. 478:343–348.
5
6
7 Brisson JA, Nuzhdin SV. 2008. Rarity of males in pea aphids results in mutational
8 decay. *Science*. 319:58.
9
10
11 Brunet S, Maro B. 2005. Cytoskeleton and cell cycle control during meiotic
12 maturation of the mouse oocyte: integrating time and space. *Reproduction*.
13 130:801–811.
14
15
16 Chauhan P, Wellenreuther M, Hansson B. 2016. Transcriptome profiling in the
17 damselfly *Ischnura elegans* identifies genes with sex-biased expression. *BMC*
18 *Genomics*. 17:985.
19
20
21 Chen Q, Yan W, Duan E. 2016. Epigenetic inheritance of acquired traits through
22 sperm RNAs and sperm RNA modifications. *Nat Rev Genet*. doi:
23 10.1038/nrg.2016.106.
24
25
26 Clark NL, Aagaard JE, Swanson WJ. 2006. Evolution of reproductive proteins
27 from animals and plants. *Reproduction*. 131:11–22.
28
29
30 Cong M et al. 2008. The enhanced immune protection of Zhikong scallop *Chlamys*
31 *farreri* on the secondary encounter with *Listonella anguillarum*. *Comp Biochem*
32 *Physiol B Biochem Mol Biol*. 151:191–196.
33
34
35 Cordero D, Delgado M, Liu B, Ruesink J, Saavedra C. 2017. Population genetics
36 of the Manila clam (*Ruditapes philippinarum*) introduced in North America and
37 Europe. *Sci Rep*. 7:39745.
38
39
40 Dadoune J-P. 2009. Spermatozoal RNAs: what about their functions? *Microsc Res*
41 *Tech*. 72:536–551.
42
43
44 Dapper AL, Wade MJ. 2016. The evolution of sperm competition genes: The effect
45 of mating system on levels of genetic variation within and between species.
46 *Evolution*. 70:502–511.
47
48
49 Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly
50 expressed proteins evolve slowly. *Proc Natl Acad Sci USA*. 102:14338–14343.
51
52
53 Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian
54 genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol*
55 *Evol*. 17:68–74.
56
57
58
59
60

- 1
2
3 Eads BD, Colbourne JK, Bohuski E, Andrews J. 2007. Profiling sex-biased gene
4 expression during parthenogenetic reproduction in *Daphnia pulex*. BMC
5 Genomics. 8:464.
6
7
8
9 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and
10 high throughput. Nucleic Acids Res. 32:1792–1797.
11
12 Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene
13 expression. Nat Rev Genet. 8:689–698.
14
15
16 Gerdol M, Venier P. 2015. An updated molecular basis for mussel immunity. Fish
17 Shellfish Immunol. 46:17–38.
18
19
20 Gershoni M, Pietrokovski S. 2014. Reduced selection and accumulation of
21 deleterious mutations in genes exclusively expressed in men. Nat Commun.
22 5:4438.
23
24
25 Ghiselli F et al. 2012. De Novo assembly of the Manila clam *Ruditapes*
26 *philippinarum* transcriptome provides new insights into expression bias,
27 mitochondrial doubly uniparental inheritance and sex determination. Mol Biol
28 Evol. 29:771–786.
29
30
31 Ghiselli F et al. 2017. The complete mitochondrial genome of the grooved carpet
32 shell, *Ruditapes decussatus* (Bivalvia, Veneridae). PeerJ. 5:e3692.
33
34
35 Goenawan IH, Bryan K, Lynn DJ. 2016. DyNet: visualization and analysis of
36 dynamic molecular interaction networks. Bioinformatics. 32:2713–2715.
37
38
39 Good JM, Nachman MW. 2005. Rates of protein evolution are positively
40 correlated with developmental timing of expression during mouse
41 spermatogenesis. Mol Biol Evol. 22:1044–1052.
42
43
44 Grath S, Parsch J. 2012. Rate of amino acid substitution is influenced by the degree
45 and conservation of male-biased transcription over 50 myr of *Drosophila*
46 evolution. Genome Biol Evol. 4:346–359.
47
48
49 Grath S, Parsch J. 2016. Sex-Biased Gene Expression. Annu Rev Genet. 50:29–44.
50
51
52 Grindstaff JL, Brodie ED 3rd, Ketterson ED. 2003. Immune function across
53 generations: integrating mechanism and evolutionary process in maternal antibody
54 transmission. Proc Roy Soc B Biol Sci. 270:2309–2319.
55
56
57 Harrison PW et al. 2015. Sexual selection drives evolution and rapid turnover of
58
59
60

1
2
3 male gene expression. Proc Natl Acad Sci USA. 112:4393–4398.

4
5
6 Hosken DJ, Hodgson DJ. 2014. Why do sperm carry RNA? Relatedness, conflict,
7 and control. Trends Ecol Evol. 29:451–455.

8
9
10 Huang G et al. 2007. The identification of lymphocyte-like cells and lymphoid-
11 related genes in amphioxus indicates the twilight for the emergence of adaptive
12 immune system. PLoS One. 2:e206.

13
14
15 Huylmans AK, López Ezquerro A, Parsch J, Cordellier M. 2016. De Novo
16 Transcriptome Assembly and Sex-Biased Gene Expression in the Cyclical
17 Parthenogenetic *Daphnia galeata*. Genome Biol Evol. 8:3120–3139.

18
19
20 Jordan IK, Mariño-Ramírez L, Wolf YI, Koonin EV. 2004. Conservation and
21 coevolution in the scale-free human gene coexpression network. Mol Biol Evol.
22 21:2058–2070.

23
24
25 Khaitovich P et al. 2005. Parallel patterns of evolution in the genomes and
26 transcriptomes of humans and chimpanzees. Science. 309:1850–1854.

27
28
29 Knorr E, Schmidtberg H, Arslan D, Bingsohn L, Vilcinskas A. 2015. Translocation
30 of bacteria from the gut to the eggs triggers maternal transgenerational immune
31 priming in *Tribolium castaneum*. Biol Lett. 11:20150885.

32
33
34 Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational
35 methods for Gene Orthology inference. Brief Bioinform. 12:379–391.

36
37
38 Kurtz J, Franz K. 2003. Innate defence: evidence for memory in invertebrate
39 immunity. Nature. 425:37–38.

40
41
42 Larracuente AM et al. 2008. Evolution of protein-coding genes in *Drosophila*.
43 Trends Genet. 24:114–123.

44
45
46 Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins
47 and gene expression levels are coupled in *Drosophila* and are independently
48 associated with mRNA abundance, protein length, and number of protein-protein
49 interactions. Mol Biol Evol. 22:1345–1354.

50
51
52 Liao B-Y, Zhang J. 2006. Evolutionary conservation of expression profiles
53 between human and mouse orthologous genes. Mol Biol Evol. 23:530–540.

54
55
56 Little TJ, O'Connor B, Colegrave N, Watt K, Read AF. 2003. Maternal transfer of
57 strain-specific immunity in an invertebrate. Curr Biol. 13:489–492.

- 1
2
3 Liu W-M et al. 2012. Sperm-borne microRNA-34c is required for the first
4 cleavage division in mouse. *Proc Natl Acad Sci USA*. 109:490–494.
5
6
7 Lotia S, Montojo J, Dong Y, Bader GD, Pico AR. 2013. Cytoscape app store.
8 *Bioinformatics*. 29:1350–1351.
9
10
11 Malone JH, Hawkins DL Jr, Michalak P. 2006. Sex-biased gene expression in a
12 ZW sex determination system. *J Mol Evol*. 63:427–436.
13
14
15 Mank JE, Ellegren H. 2009. Are sex-biased genes more dispensable? *Biol Lett*.
16 5:409–412.
17
18
19 Mank JE, Hultin-Rosenberg L, Axelsson E, Ellegren H. 2007. Rapid evolution of
20 female-biased, but not male-biased, genes expressed in the avian brain. *Mol Biol*
21 *Evol*. 24:2698–2706.
22
23
24 Mank JE, Hultin-Rosenberg L, Zwahlen M, Ellegren H. 2008. Pleiotropic
25 constraint hampers the resolution of sexual antagonism in vertebrate gene
26 expression. *Am Nat*. 171:35–43.
27
28
29 Mank JE, Nam K, Brunström B, Ellegren H. 2010. Ontogenetic complexity of
30 sexual dimorphism and sex-specific selection. *Mol Biol Evol*. 27:1570–1578.
31
32
33 Marlow FL. 2011. *Maternal Control of Development in Vertebrates: My Mother*
34 *Made Me Do It!* Morgan & Claypool Life Sciences: San Rafael (CA).
35
36
37 Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-
38 biased gene expression in *Drosophila*. *Proc Natl Acad Sci USA*. 100:9894–9899.
39
40
41 Metta M, Gudavalli R, Gibert J-M, Schlötterer C. 2006. No accelerated rate of
42 protein evolution in male-biased *Drosophila pseudoobscura* genes. *Genetics*.
43 174:411–420.
44
45
46 Milani L, Ghiselli F, Nuzhdin SV, Passamonti M. 2013. Nuclear genes with sex
47 bias in *Ruditapes philippinarum* (Bivalvia, veneridae): Mitochondrial inheritance
48 and sex determination in DUI species. *J Exp Zool B Mol Dev Evol*. 320:442–454.
49
50
51 Miller D, Ostermeier GC, Krawetz SA. 2005. The controversy, potential and roles
52 of spermatozoal RNA. *Trends Mol Med*. 11:156–163.
53
54
55 Milutinović B, Kurtz J. 2016. Immune memory in invertebrates. *Semin Immunol*.
56 28:328–342.
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Moreira R et al. 2012. Gene expression analysis of clams *Ruditapes philippinarum* and *Ruditapes decussatus* following bacterial infection yields molecular insights into pathogen resistance and immunity. *Dev Comp Immunol.* 36:140–149.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 5:621–628.

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics.* doi: 10.1093/bioinformatics/btx445.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol.* 21:1308–1317.

Papa F et al. 2017. Rapid evolution of female-biased genes among four species of *Anopheles malaria* mosquitoes. *Genome Res.* doi: 10.1101/gr.217216.116.

Parisi M et al. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science.* 299:697–700.

Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nat Rev Genet.* 14:83–87.

Pauletto M et al. 2014. Deep transcriptome sequencing of *Pecten maximus* hemocytes: a genomic resource for bivalve immunology. *Fish Shellfish Immunol.* 37:154–165.

Poley JD, Sutherland BJG, Jones SRM, Koop BF, Fast MD. 2016. Sex-biased gene expression and sequence conservation in Atlantic and Pacific salmon lice (*Lepeophtheirus salmonis*). *BMC Genomics.* 17:483.

Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics.* 174:893–900.

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-Dependent Gene Expression and Evolution of the *Drosophila* Transcriptome. *Science.* 300:1742–1745.

Reinke V, Gil IS, Ward S, Kazmer K. 2004. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development.* 131:311–323.

- 1
2
3
4 Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology
5 Open Software Suite. *Trends Genet.* 16:276–277.
6
7
8 Rice WR. 2013. Nothing in Genetics Makes Sense Except in Light of Genomic
9 Conflict. *Annu Rev Ecol Evol Syst.* 44:217–237.
10
11 Rice WR, Chippindale AK. 2001. Intersexual ontogenetic conflict. *J Evol Biol.*
12 14:685–693.
13
14
15 Ridout KE, Dixon CJ, Filatov DA. 2010. Positive selection differs between protein
16 secondary structure elements in *Drosophila*. *Genome Biol Evol.* 2:166–179.
17
18
19 Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression
20 and the evolution of gene regulation. *Nat Rev Genet.* 13:505–516.
21
22
23 Sartor MA et al. 2006. A new method to remove hybridization bias for interspecies
24 comparison of global gene expression profiles uncovers an association between
25 mRNA sequence divergence and differential gene expression in *Xenopus*. *Nucleic*
26 *Acids Res.* 34:185–200.
27
28
29 Schulenburg H, Boehnisch C, Michiels NK. 2007. How do invertebrates generate a
30 highly specific innate immune response? *Mol Immunol.* 44:3338–3344.
31
32
33 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015.
34 BUSCO: assessing genome assembly and annotation completeness with single-
35 copy orthologs. *Bioinformatics.* 31:3210–3212.
36
37
38 Small CM, Carney GE, Mo Q, Vannucci M, Jones AG. 2009. A microarray
39 analysis of sex- and gonad-biased gene expression in the zebrafish: Evidence for
40 masculinization of the transcriptome. *BMC Genomics.* 10:579.
41
42
43 Sperry AO. 2012. The dynamic cytoskeleton of the developing male germ cell.
44 *Biol Cell.* 104:297–305.
45
46
47 Stewart AD, Pischedda A, Rice WR. 2010. Resolving intralocus sexual conflict:
48 genetic mechanisms and time frame. *J Hered.* 101 Suppl 1:S94–9.
49
50
51 Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and
52 visualizes long lists of gene ontology terms. *PLoS One.* 6:e21800.
53
54
55 Swanson WJ, Vacquier VD. 2002. The Rapid Evolution of Reproductive Proteins.
56 *Nat Rev Genet.* 3:137–144.
57
58
59
60

- 1
2
3 Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed
4 sequence tag analysis of *Drosophila* female reproductive tracts identifies genes
5 subjected to positive selection. *Genetics*. 168:1457–1465.
6
7
8
9 Tirosh I, Barkai N. 2008. Evolution of gene sequence and gene expression are not
10 correlated in yeast. *Trends Genet*. 24:109–113.
11
12 Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are
13 rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol*
14 *Biol Evol*. 19:1973–1980.
15
16
17 Trachana K et al. 2011. Orthology prediction methods: a quality assessment using
18 curated protein families. *Bioessays*. 33:769–780.
19
20
21 Turner LM, Chuong EB, Hoekstra HE. 2008. Comparative analysis of testis
22 protein evolution in rodents. *Genetics*. 179:2075–2089.
23
24
25 Turner LM, Hoekstra HE. 2008. Causes and consequences of the evolution of
26 reproductive proteins. *Int J Dev Biol*. 52:769–780.
27
28
29 Voolstra CR, GIGA Community of Scientists (COS), Wörheide G, Lopez JV.
30 2017. Advancing Genomics through the Global Invertebrate Genomics Alliance
31 (GIGA). *Invertebr Syst*. 31:1–7.
32
33
34 Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit
35 incorporating gamma-series methods and sliding window strategies. *Genomics*
36 *Proteomics Bioinformatics*. 8:77–80.
37
38
39 Wang L, Yue F, Song X, Song L. 2015. Maternal immune transfer in mollusc. *Dev*
40 *Comp Immunol*. 48:354–359.
41
42
43 Wang X, Werren JH, Clark AG. 2015. Genetic and epigenetic architecture of sex-
44 biased expression in the jewel wasps *Nasonia vitripennis* and *giraulti*. *Proc Natl*
45 *Acad Sci USA*. 112:E3545–54.
46
47
48 Wang Y, Coleman-Derr D, Chen G, Gu YQ. 2015. OrthoVenn: a web server for
49 genome wide comparison and annotation of orthologous clusters across multiple
50 species. *Nucleic Acids Res*. 43:W78–84.
51
52
53 Watson AJ. 2007. Oocyte cytoplasmic maturation: a key mediator of oocyte and
54 embryo developmental competence. *J Anim Sci*. 85:E1–3.
55
56
57 Yang L, Zhang Z, He S. 2016. Both Male-Biased and Female-Biased Genes
58
59
60

1
2
3 Evolve Faster in Fish Genomes. *Genome Biol Evol.* 8:3433–3445.

4
5
6 Yang X et al. 2006. Tissue-specific expression and regulation of sexually
7 dimorphic genes in mice. *Genome Res.* 16:995–1004.

8
9
10 Yuan S et al. 2015. mir-34b/c and mir-449a/b/c are required for spermatogenesis,
11 but not for the first cleavage division in mice. *Biol Open.* 4:212–223.

12
13 Yuan S, Tao X, Huang S, Chen S, Xu A. 2014. Comparative immune systems in
14 animals. *Annu Rev Anim Biosci.* 2:235–258.

15
16
17 Yue F et al. 2013. Maternal transfer of immunity in scallop *Chlamys farreri* and its
18 trans-generational immune protection to offspring against bacterial challenge. *Dev*
19 *Comp Immunol.* 41:569–577.

20
21
22 Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution.
23 *Nat Rev Genet.* 16:409–420.

24
25
26 Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B. 2007. Constraint and turnover
27 in sex-biased gene expression in the genus *Drosophila*. *Nature.* 450:233–237.

28
29
30 Zhang Z, Hambuch TM, Parsch J. 2004. Molecular evolution of sex-biased genes
31 in *Drosophila*. *Mol Biol Evol.* 21:2130–2139.

32
33
34 Zhang Z, Parsch J. 2005. Positive correlation between evolutionary rate and
35 recombination rate in *Drosophila* genes with male-biased expression. *Mol Biol*
36 *Evol.* 22:1945–1947.

37
38
39 Zouros E. 2013. Biparental Inheritance Through Uniparental Transmission: The
40 Doubly Uniparental Inheritance (DUI) of Mitochondrial DNA. *Evol Biol.* 40:1–31.

Figure Legends

Figure 1

Volcano plot of the transcription in *R. decussatus* (top) and *R. philippinarum* (bottom). Male-enriched transcripts are represented in blue, female-enriched transcripts in red, unbiased transcripts in grey. Dashed lines: $\log_2(\text{fold change}) = -1, 1$ (fold change = -2, 2).

Figure 2

Sex bias difference in groups of orthologs between *R. decussatus* and *R. philippinarum* (N=1521), as represented by a cluster analysis performed using the SCALE index. Cluster B: no difference in transcription bias between groups of orthologs; yellow = *R. philippinarum*; green = *R. decussatus*.

Figure 3

Kernel density plot of the distribution of dN/dS in unbiased genes in both the species (green line), in genes that are unbiased in one species and male- or female-biased in the other species (respectively blue line, and pink line), in genes where the sex bias is maintained (black line for male-biased genes, red line for female-biased genes) and in genes with a reversed sex bias (yellow line).

Figure 4

a) Relationship between dN/dS and amino acid p-distance of all orthologous genes between *R. decussatus* and *R. philippinarum*. A linear function (black dashed line) describes the relationship between dN/dS and p-distance in genes with lower p-distance. In genes with p-distance higher than 40%, the relationship is better explained by an exponential function (red dashed line). b) unbiased genes in both the species (green); c) genes that are unbiased in one species and male biased in the other (blue); d) genes that are unbiased in one species and female-biased in the other species (pink); e) genes where a male-bias is maintained (black); f) genes where a female-bias is maintained (red). Dashed lines in b-f represent the regression lines corresponding to the linear model calculated for all genes; solid

1
2
3 colored lines in b-f represent the regression lines corresponding to the linear
4 models calculated with the specific subset of genes.
5
6
7

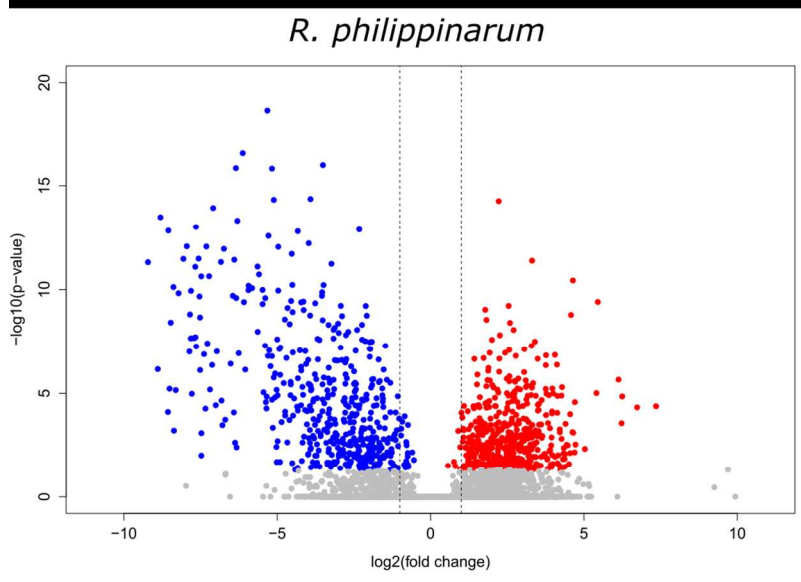
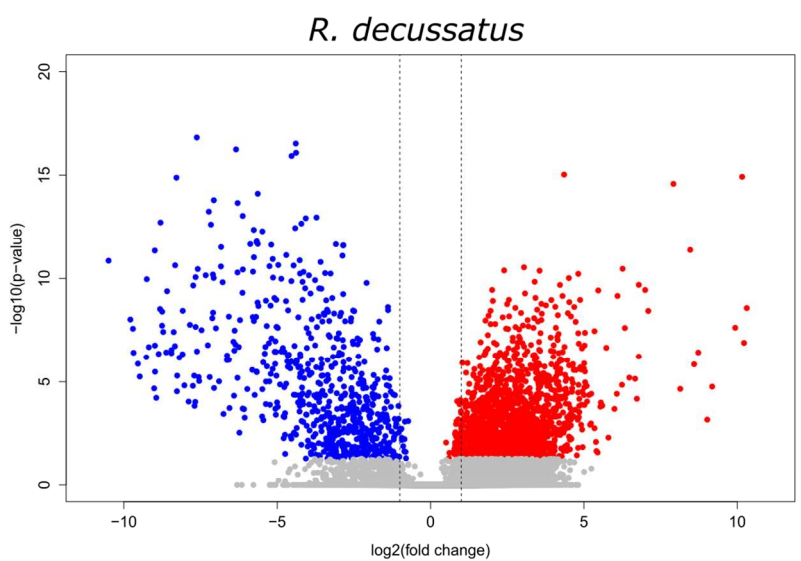
8 Figure 5
9

10 Plot indicating the relationship between the rate of protein sequence evolution
11 indicated as $\log_2(dN/dS)$, and transcription level indicated as $\log_2(FPKM)$ in *R.*
12 *decussatus* (left), and *R. philippinarum* (right).
13
14
15

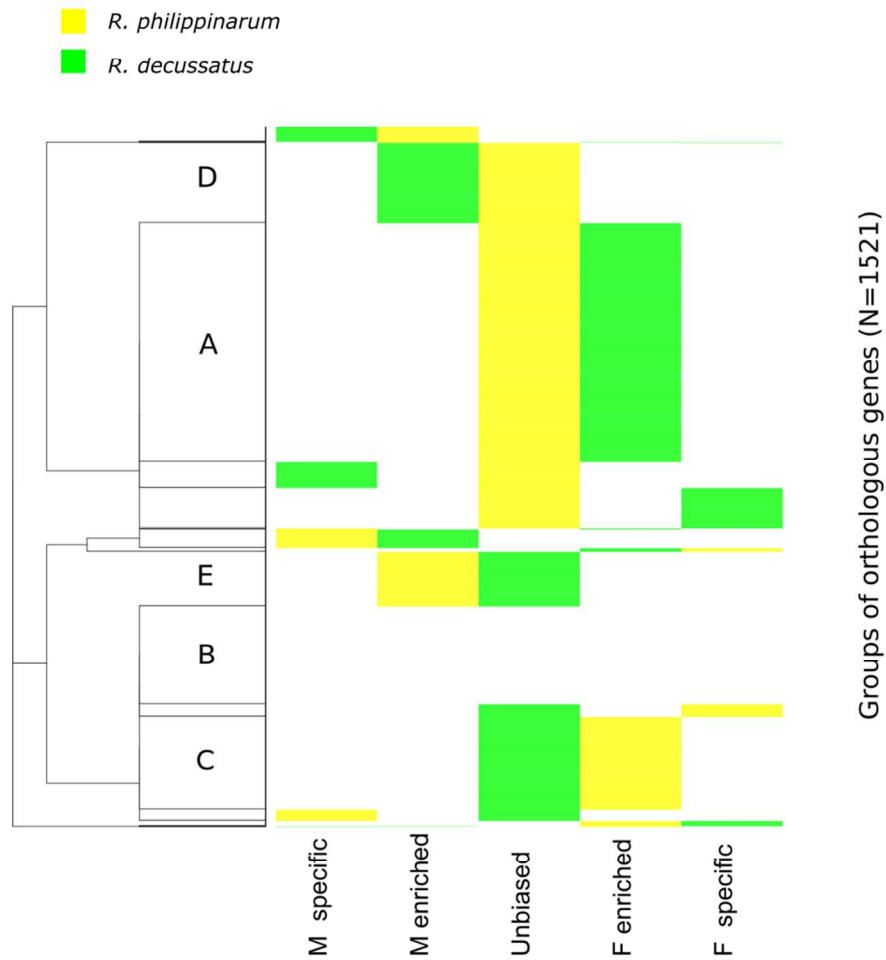
16 Figure 6
17

18 Network of GO term enrichment in singlets of *R. decussatus* (green), and *R.*
19 *philippinarum* (red).
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

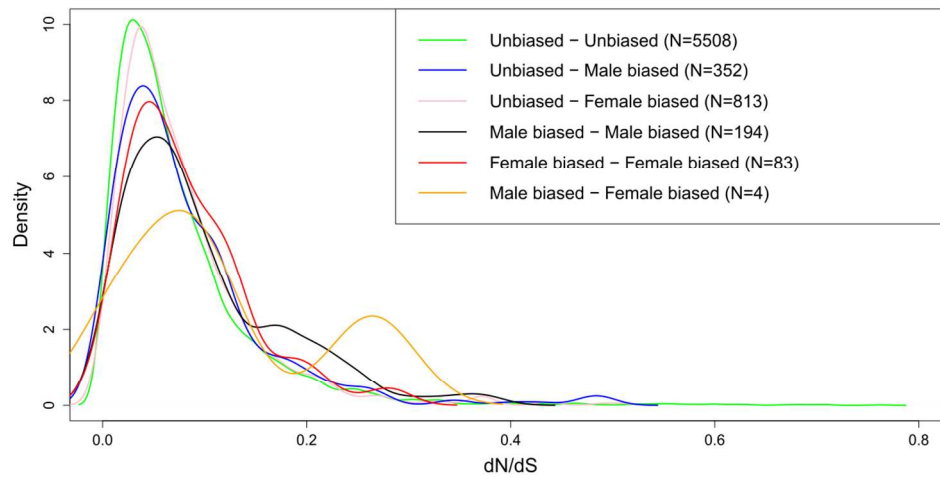


Volcano plot of the transcription in *R. decussatus* (top) and *R. philippinarum* (bottom). Male-enriched transcripts are represented in blue, female-enriched transcripts in red, unbiased transcripts in grey. Dashed lines: $\log_2(\text{fold change}) = -1, 1$ (fold change = -2, 2).



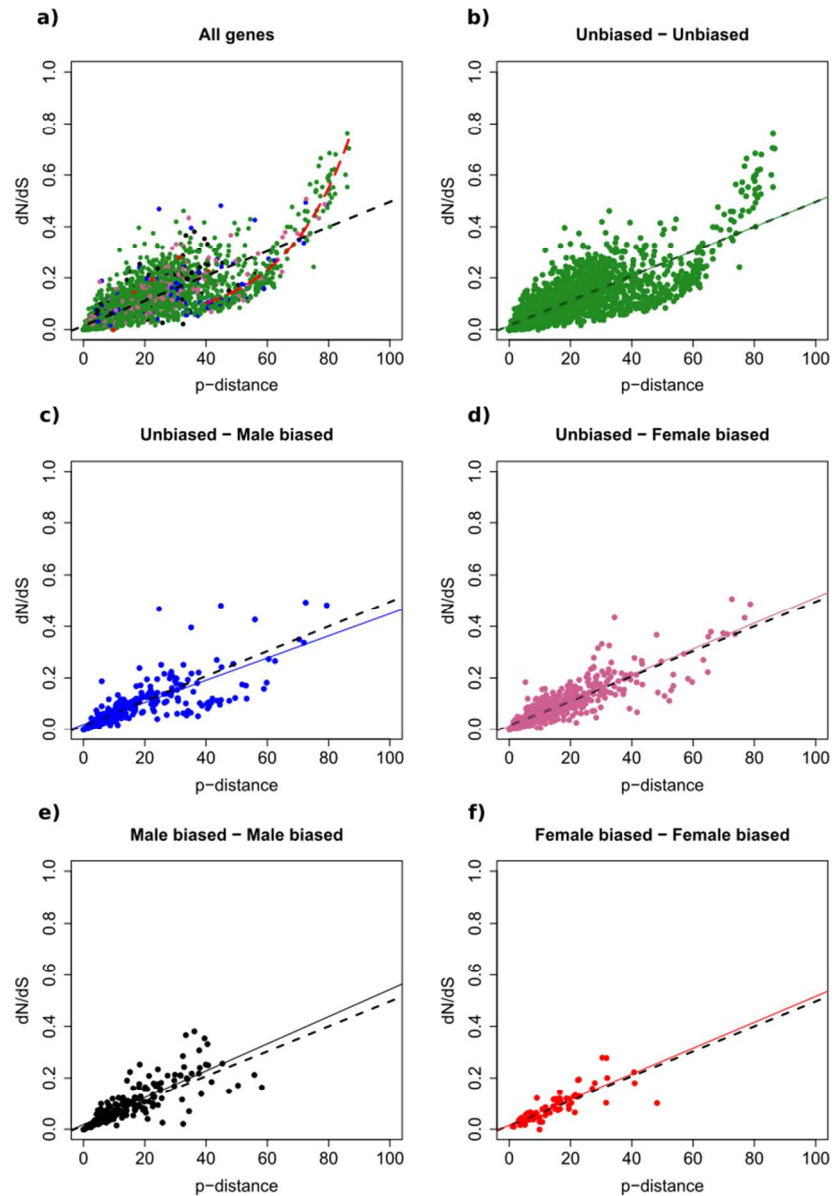
Sex bias difference in groups of orthologs between *R. decussatus* and *R. philippinarum* (N=1521), as represented by a cluster analysis performed using the SCALE index. Cluster B: no difference in transcription bias between groups of orthologs; yellow=*R. philippinarum*; green=*R. decussatus*.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



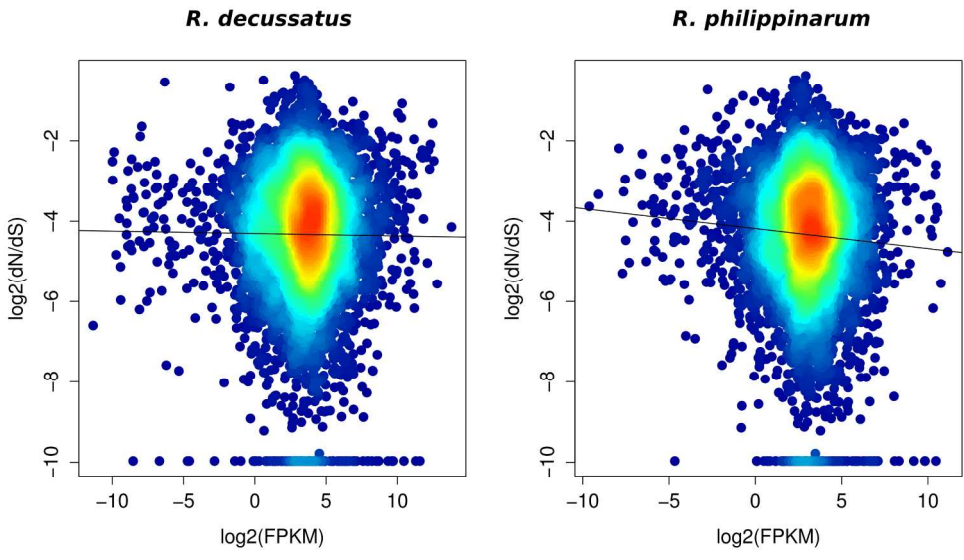
Kernel density plot of the distribution of dN/dS in unbiased genes in both the species (green line), in genes that are unbiased in one species and male- or female-biased in the other species (respectively blue line, and pink line), in genes where the sex bias is maintained (black line for male-biased genes, red line for female-biased genes) and in genes with a reversed sex bias (yellow line).

er Review



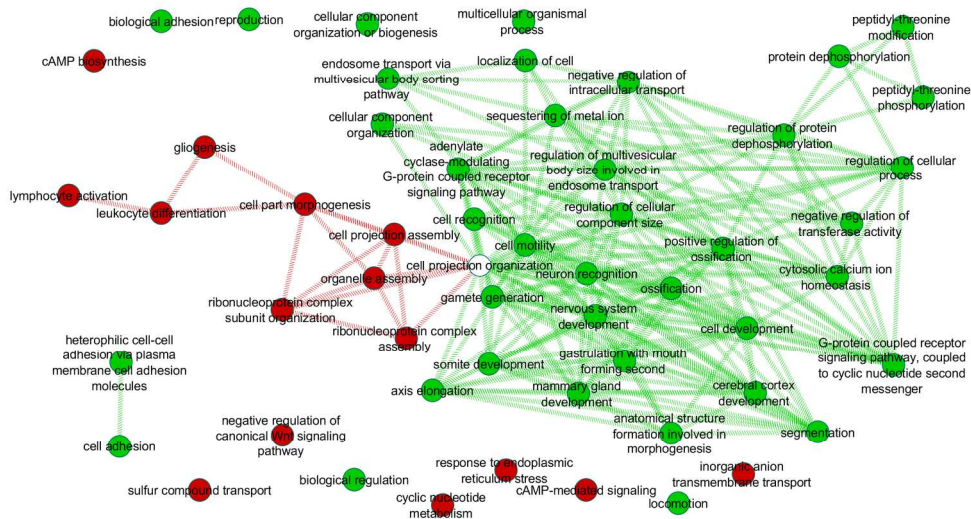
a) Relationship between dN/dS and amino acid p-distance of all orthologous genes between *R. decussatus* and *R. philippinarum*. A linear function (black dashed line) describes the relationship between dN/dS and p-distance in genes with lower p-distance. In genes with p-distance higher than 40%, the relationship is better explained by an exponential function (red dashed line). b) unbiased genes in both the species (green); c) genes that are unbiased in one species and male biased in the other (blue); d) genes that are unbiased in one species and female-biased in the other species (pink); e) genes where a male-bias is maintained (black); f) genes where a female-bias is maintained (red). Dashed lines in b-f represent the regression lines corresponding to the linear model calculated for all genes; solid colored lines in b-f represent the regression lines corresponding to the linear models calculated with the specific subset of genes.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



Plot indicating the relationship between the rate of protein sequence evolution indicated as $\log_2(dN/dS)$, and transcription level indicated as $\log_2(FPKM)$ in *R. decussatus* (left), and *R. philippinarum* (right).

Peer Review



Network of GO term enrichment in singlets of *R. decussatus* (green), and *R. philippinarum* (red).

Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1: Number (and percentage) of sex-biased transcripts in *R. decussatus* and *R. philippinarum*.

	M-specific	M-enriched	Unbiased	F-enriched	F-specific
<i>R. decussatus</i>	775 (2%)	661 (1.7%)	35,532 (90%)	1,589 (4%)	910 (2.3%)
<i>R. philippinarum</i>	448 (2%)	909 (4%)	20,772 (90.7%)	631 (2.8%)	126 (0.5%)

For Peer Review

Table 2: Annotation statistics of *R. decussatus* and *R. philippinarum* transcriptomes.

	Assembled Loci	Protein Annotation	GO Annotation	Nucleotide Annotation	Total Annotated	Not Annotated
<i>R. decussatus</i>	39,467	14,315 (36.3%)	13,865 (35.1%)	13,697 (34.7%)	28,022 (71%)	11,445 (29%)
<i>R. philippinarum</i>	22,886	12,371 (54%)	12,064 (52.7%)	3,997 (17.5%)	16,436 (71.8%)	6,450 (28.2%)

For Peer Review

Table 3: comparison of transcription sex bias of orthologous groups between *R. decussatus* and *R. philippinarum* according to the SCALE index.

Sex Bias of Orthologous Groups		Number of Orthologous Groups (%)	Cluster (Fig.2)
<i>R. decussatus</i>	<i>R. philippinarum</i>		
Female Enriched	Unbiased	521 (34.2%)	A
	Sex Bias Maintained	213 (14%)	B
Unbiased	Female Enriched	201 (13.2%)	C
Male Enriched	Unbiased	175 (11.5%)	D
Unbiased	Male Enriched	120 (7.8%)	E
Female Specific	Unbiased	87 (5.7%)	na
Male Specific	Unbiased	56 (3.7%)	na
Male Enriched	Male Specific	40 (2.6%)	na
Male Specific	Male Enriched	32 (2.1%)	na
Unbiased	Female Specific	27 (1.8%)	na
Unbiased	Male Specific	24 (1.6%)	na
Female Specific	Female Enriched	11 (0.7%)	na
Female Enriched	Female Specific	8 (0.5%)	na
Female Enriched	Male Specific	2 (0.1%)	na
Male Specific	Female Enriched	1 (0.07%)	na
Male Enriched	Female Enriched	1 (0.07%)	na
Female Enriched	Male Enriched	1 (0.07%)	na
Female Specific	Male Enriched	1 (0.07%)	na
Male Specific	Female Specific	0	na

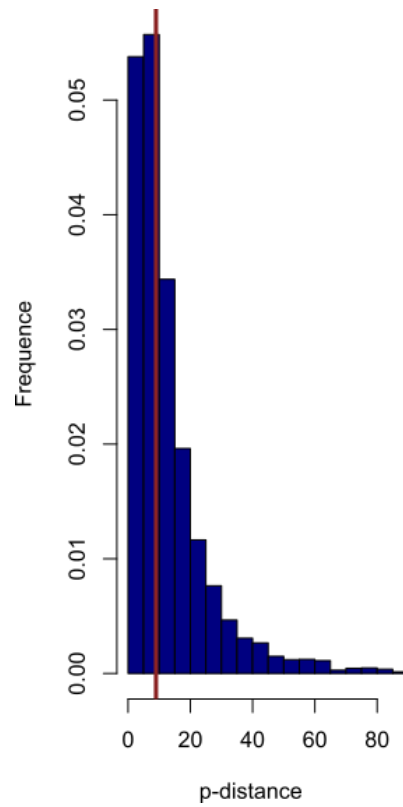
Table 4: top 20 most frequent GO terms associated with 'fast mutating' ($dN/dS < 0.2$, $40\% \leq p\text{-distance} \leq 60\%$), and 'fast evolving' orthologs ($dN/dS > 0.2$, $p\text{-distance} > 60\%$). Of the 25 GO terms represented in this list, 15 appear in both groups of orthologs, while 10 (underlined) appear only in one group.

'Fast-mutating' orthologs

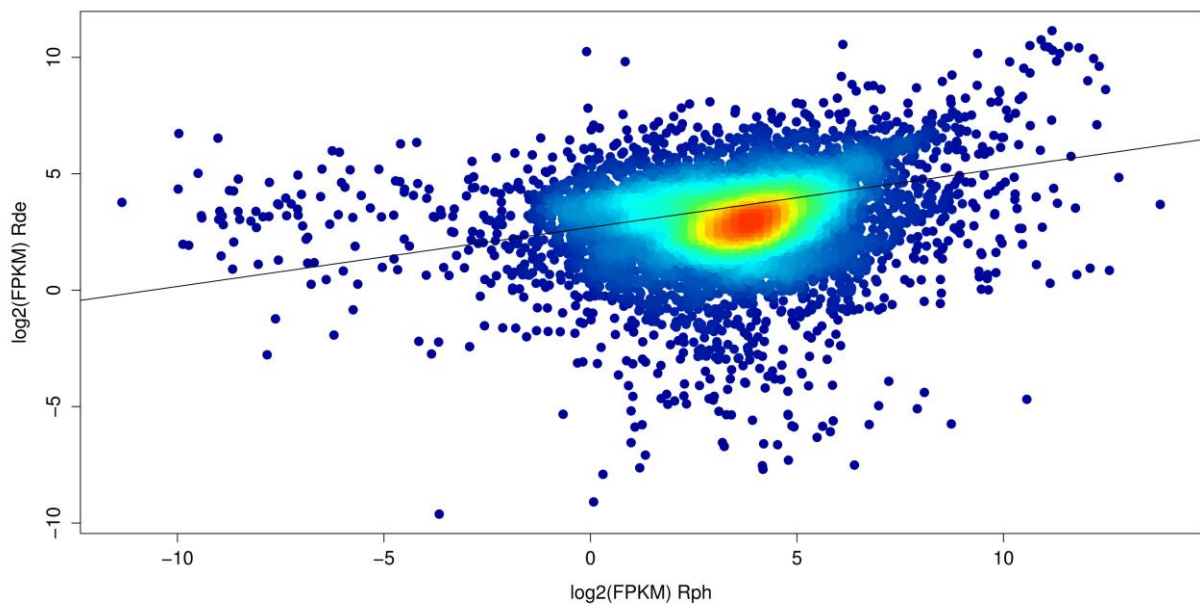
'Fast-evolving' orthologs

metabolic process	metabolic process
<u>biosynthetic process</u>	<u>nucleobase-containing compound metabolic process</u>
transport	transport
oxidation-reduction process	oxidation-reduction process
<u>response to stimulus</u>	transcription, DNA-templated
transcription, DNA-templated	regulation of transcription, DNA-templated
<u>regulation of RNA biosynthetic process</u>	transmembrane transport
<u>regulation of nucleic acid-templated transcription</u>	phosphorylation
regulation of transcription, DNA-templated	cellular protein modification process
transmembrane transport	signal transduction
phosphorylation	translation
cellular protein modification process	ion transport
signal transduction	carbohydrate metabolic process
translation	proteolysis
ion transport	protein phosphorylation
carbohydrate metabolic process	intracellular signal transduction
proteolysis	<u>lipid metabolic process</u>
protein phosphorylation	<u>nucleotide biosynthetic process</u>
intracellular signal transduction	<u>cellular response to DNA damage stimulus</u>
<u>ion transmembrane transport</u>	<u>nucleic acid phosphodiester bond hydrolysis</u>

Supplementary figures and tables



Supplementary figure 1. Distribution of amino acid p-distance of orthologous genes between *R. decussatus* and *R. philippinarum*.



Supplementary figure 2. Transcription levels correlation between orthologous genes of *R. decussatus* (log₂(FPKM) Rde) and *R. philippinarum* (log₂(FPKM) Rph).

Supplementary Table 1: *R. decussatus* de novo assembly statistics.

Total number of reads	67,386,984
Total nucleotides	5,121,410,784
Contigs	
Total number of contigs	69,279
Median length of contig sequences	795
N50 length of contig sequences	2,064
Representative loci sequences	
Total number of loci	39,467
Median length of loci sequences	668
N50 length of loci sequences	1,672
Total length of all loci sequences	43,791,854
Number of loci having multiple contigs	12,300

Supplementary Table 2: completeness assessment of transcriptome assemblies according to BUSCO v2 as implemented in gVolante (<https://gvolante.riken.jp/>).
eBUSCO = Eukaryota ortholog set; mBUSCO = Metazoa ortholog set.

	<i>R. decussatus</i>		<i>R. philippinarum</i>	
	eBUSCO	mBUSCO	eBUSCO	mBUSCO
Total # of core genes queried	303	978	303	978
Complete	276 (91.09%)	853 (87.22%)	224 (73.93%)	692 (70.76%)
Complete + Partial	292 (96.37%)	926 (94.68%)	282 (93.07%)	859 (87.83%)
Missing	11 (3.63%)	52 (5.32%)	21 (6.93%)	119 (12.17%)
SEQUENCE STATISTICS				
# of sequences	39,467		22,818	
Total length (nt)	43,791,854		17,963,136	
Longest sequence (nt)	28,079		24,386	
Shortest sequence (nt)	300		259	
Mean sequence length (nt)	1,110		787	
Median sequence length (nt)	668		502	
N50 sequence length (nt)	1,672		1,006	
L50 sequence length (nt)	7,117		4,649	
# of sequences > 1K (nt)	13,276 (33.6%)		4,687 (20.5%)	
# of sequences > 10K (nt)	60 (0.2%)		13 (0.1%)	
GC-content (%)	35.05		39.18	
Base composition (%)	A:33.08, C:17.16, G:17.77, T:31.66, N:0.33		A:32.27, C:18.76, G:20.41, T:28.55, N:0.01	

Supplementary Table 3: BLAST annotation of the orthologous groups that showed an opposite transcription sex bias between *R. decussatus* and *R. philippinarum*.

Sex Bias of Orthologous Groups		BLAST Annotation
<i>R. decussatus</i>	<i>R. philippinarum</i>	
Male Specific	Female Enriched	Pre-mRNA-processing factor 40-like A
Male Enriched	Female Enriched	Coiled-coil-helix-coiled-coil-helix domain containing protein 3; Mitochondrial
Female Specific	Male Enriched	Probable tRNA N6-adenosine threonylcarbamoyltransferase; Mitochondrial
Female Enriched	Male Specific	Innexin
Female Enriched	Male Specific	E3 ubiquitin-protein ligase MARCH2
Female Enriched	Male Specific	ATP-binding cassette sub-family B member 10; Mitochondrial

Supplementary Materials 1-5 are available on figshare: <https://figshare.com/s/55515385c449ca16c042>

Chapter 4:

No evidence for nuclear compensation hypothesis in species with different mechanisms of mitochondrial inheritance

Mariangela Iannello¹, Guglielmo Puccio¹, Federico Plazzi¹, Fabrizio Ghiselli^{1,*†}, and Marco Passamonti^{1,†}

Affiliations:

¹Department of Biological, Geological, and Environmental Sciences – University of Bologna, Italy

[†] These authors contributed equally to this work.

*Author for Correspondence: Fabrizio Ghiselli, Department of Biological, Geological, and Environmental Sciences, University of Bologna, Bologna, Italy, telephone: +39 051 2094203, email address: fabrizio.ghiselli@unibo.it

Abstract

Mitochondria are a fundamental component of the eukaryotic cell that derived from a free living α -proteobacterium. Although mitochondria retained part of their original genome, most of the genes involved in biogenesis, transmission and homeostasis of these organelles are encoded by the nucleus. Also, the genes involved in the main process of energy production of most eukaryotes (oxidative phosphorylation, OXPHOS) are encoded by both nuclear (nuDNA) and mitochondrial (mtDNA) genomes. Therefore a tight co-regulation between mtDNA and nuDNA is essential for mitochondrial activity, and it is the result of a long-lasting mitonuclear co-evolution. It is commonly accepted that Metazoa have evolved a non-mendelian mechanism of mitochondrial inheritance, to avoid the presence of mixed mtDNA haplotypes in the same organism and reduce genomic conflicts and mito-nuclear incompatibilities. Several works investigated the effects of having different mtDNA variants working with the same nuclear background by producing cytoplasmic hybrids. These works show detrimental effects of heteroplasmy, such as reduction of OXPHOS activity, oxidative damage, disruption of mitochondrial functions.

While most eukaryotes are characterized by a Strictly Maternal Inheritance (SMI) of mitochondria, some species of bivalve molluscs present the Doubly Uniparental inheritance (DUI), where two distinct mitochondrial lineages are present: F-type, inherited through eggs, and M-type, inherited through sperm. So, in DUI species, the same nuclear background have to coordinate two different mitochondrial mtDNAs, characterized by a high nucleotide divergence (20-40%) and different replication and transcription dynamics.

In this work we took advantage of the natural heteroplasmic condition of DUI species to get insights into the dynamics of mito-nuclear co-evolution. Particularly, we used RNA-Seq to investigate the transcription and rate of protein evolution of OXPHOS genes in gonads of two related bivalve species: *Ruditapes decussatus*, a species with SMI, and *Ruditapes philippinarum*, a species with DUI.

Introduction

Mitochondria are a primary component of the eukaryotic cell and they are characterized by peculiar features concerning their origin, evolution, and inheritance mechanism. It is now commonly acknowledged that mitochondria derived from a free-living α -proteobacterium that ~2 billion years ago became an endosymbiont of an archaeon (see Martin et al. 2015 for a thorough review), and this event played a crucial role in the evolution of eukaryotes (see for example: Martin & Koonin 2006; Lane & Martin 2010). Since then, a massive gene transfer from the ancestral mitochondrial genome to the nucleus (endosymbiotic gene transfer) occurred during evolution, and most of the genes involved in mitochondrial function and biogenesis are now encoded by the nucleus (Timmis et al. 2004). Nevertheless, mitochondria retained part of their original genome, and the proteins involved in the main process of energy production of most eukaryotes—that is oxidative phosphorylation (OXPHOS)—are encoded by either the nuclear (nuDNA) or the mitochondrial genome (mtDNA), and they have to function together in 4 out of 5 complexes (Complex I, III, IV, and ATPase). Accordingly, it was proposed that a tight co-evolution and co-regulation of these two genomes is essential to maintain an efficient mitochondrial activity (see for example: Rand et al. 2004; Bar-Yaacov et al. 2012; Allen 2015). Still, mitochondrial OXPHOS subunits seem to be subject to different evolutionary forces compared to nuclear subunits. For instance, the rate of amino acid sequence evolution (calculated as the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site, dN/dS) of mitochondrial subunits is remarkably lower than that of nuclear subunits in all taxa investigated so far (see for example: Nabholz et al. 2012; Popadin et al. 2012). Furthermore, there is evidence for strong purifying selection acting on mitochondrially-encoded OXPHOS subunits (Popadin et al. 2012; Piganeau & Eyre-Walker 2009), even if signs of positive selection were also reported (James et al. 2016; Castellana et al. 2011; Pavlova et al. 2017; Gibson et al. 2010). Three hypotheses were proposed to explain such differences: first, mitochondrial OXPHOS subunits could be subject to tighter functional constraints, since they assemble the core of OXPHOS complexes, while nuclear subunits—that are instead more peripheral—could be under a more relaxed selection (Zhang & Broughton 2013; Popadin et al. 2012). Second, as mitochondrial genome in animal tends to accumulate mutations from 9 to 25 times faster than the nuclear genome (Lynch et al. 2006), positive selection would act on nuclear OXPHOS subunits to compensate the insurgence of mutations in mitochondrial genes, ensuring a proper structural/functional match among the subunits of OXPHOS complexes. This theory, called ‘nuclear compensation hypothesis’, was adopted by several authors (Burton et al. 2006; Dowling et al. 2008; Osada & Akashi 2012; Havird & Sloan 2016; Aanen et al. 2014; Burton & Barreto 2012). Third, Nabholz et al. (2012) proposed that transcription level could be the main responsible in affecting the rate of protein evolution in OXPHOS genes, supporting the assumption that a negative correlation—called ‘E-R

correlation'—exists between transcripts abundance and rate of protein evolution (Zhang & Yang 2015). How tight is the co-evolution between mitochondrial and nuclear genomes? So far many works investigated the effects of having different mtDNA variants working in the same nuclear background by producing cytoplasmic hybrids. Some of these experiments included hybrids carrying mitochondrial and nuclear genome of different species (McKenzie et al. 2003; Sackton et al. 2003; Niehuis et al. 2008), showing good evidence for detrimental effects of heteroplasmy, such as reduction of OXPHOS activity, oxidative damage, and disruption of mitochondrial functions (see for example: Kazuno et al. 2006; Moreno-Loshuertos et al. 2006; Barreto & Burton 2013). Similar results were obtained when inter-population hybrids carried mitochondrial variants different from those co-adapted with their respective nuclear background (Barreto & Burton 2013; Ellison & Burton 2008; Sharpley et al. 2012). Therefore, cytonuclear incompatibilities are also believed to play an important role in postzygotic barriers, in adaptation and speciation (Hill 2016, 2017; Ballard et al. 2007; Wolff et al. 2014; Gershoni et al. 2009).

Since mixing different mitochondrial lineages can foster the emergence of genomic conflicts or generate mito-nuclear incompatibilities, it has been proposed that Metazoa have evolved a non-mendelian mechanism of mitochondrial inheritance, in order to avoid the presence of several mtDNA haplotypes in the same organism (see for example Birky 1995; Lane 2012). Indeed, metazoans are almost invariably characterized by SMI of mitochondria (Birky 1995), namely only females transmit mitochondria to the offspring, while paternal mitochondrial contribution is avoided in very different ways across eukaryotes (Birky 1995; Sato & Sato 2013). Until now, the only known evolutionarily stable exception to SMI in Metazoa is the Doubly Uniparental Inheritance (DUI; Skibinski et al. 1994a, 1994b; Zouros, Ball, et al. 1994; Zouros, Oberhauser Ball, et al. 1994), a peculiar mechanism of mitochondrial heredity observed in ~100 species of bivalve molluscs (Gusman et al. 2016). In DUI species, two distinct mitochondrial lineages are present: the F-type, inherited through eggs, and the M-type, inherited through sperm. While F-type mtDNA is transmitted from females to all the progeny, M-type mtDNA is transmitted from males to male progeny only. Consequently, females are homoplasmic for the F-type, while males are heteroplasmic. Therefore, differently to all other Metazoa, in DUI bivalves the same nuclear background had to co-evolve with two DISTINCT mtDNAs, presenting a high nucleotide divergence (up to 40%; Zouros 2013) and different replication and transcription dynamics (Ghiselli et al. 2011; Obata et al. 2011; Ghiselli et al. 2013; Milani et al. 2014; Guerra et al. 2016). Thus, DUI species offer a unique opportunity to investigate the effects of heteroplasmy, without the need to generate cytoplasmic hybrids: in the DUI male, heteroplasmy is natural, therefore its biological functions and interactions between nucleus and mitochondria are the unaltered result of evolution.

In this work we investigated the rate of protein evolution and the transcription level of nuclear and mtDNA-encoded OXPHOS subunits in two related bivalve species: *Ruditapes decussatus*, characterized by SMI of mitochondria, and *Ruditapes philippinarum*, with DUI of mitochondria. We also examined the relationship between transcription level and dN/dS in OXPHOS subunits.

Lastly, we investigated the rate of protein evolution of nuclear genes involved in mitochondrial functions.

Materials and Methods

Data

Transcriptome data of mature gonads from twelve individuals (six females and six males) of *R. decussatus* and twelve individuals (six females and six males) of *R. philippinarum* were retrieved respectively from Ghiselli et al. (submitted to Genome Biology and Evolution, pending major revisions) and Ghiselli et al. (2012). Raw reads and transcriptome assembly from both experiments are available on NCBI, under BioProjects PRJNA68513 (*R. philippinarum*) and PRJNA170478 (*R. decussatus*). For technical details about the *de novo* assembly and differential transcription analysis between males and females refer to Ghiselli et al. (2012). Both transcriptomes were annotated using a pipeline specifically designed for non-model organisms (Ghiselli et al. in preparation; protocol and detailed information available here: https://osf.io/cdkb9/?view_only=f0b2cde926db43719f3d705012c4eeaa). Orthologous genes between the two species were found using OrthoVenn (Wang et al. 2015) with default parameters. Nuclear-encoded OXPHOS subunits were found combining the protein annotation obtained by BLASTP (Camacho et al. 2009), HMMER3 (Mistry et al. 2013) and InterProScan 5.3-46.0 (Jones et al. 2014), as implemented in the above-mentioned annotation pipeline. MtDNA-encoded OXPHOS subunits were retrieved from the mitochondrial genomes available on NCBI GenBank under the Accession Numbers AB065374 and AB065375 for *R. philippinarum* M and F mtDNAs, and KP089983 for *R. decussatus* (Ghiselli et al. 2017). Transcription level of mtDNA-encoded OXPHOS subunits was calculated by mapping the reads on whole mitochondrial genomes with BWA (Li & Durbin 2009).

Transcription of OXPHOS genes

We performed a comparative transcription analysis of nuclear- and mtDNA-encoded subunits of complexes involved in OXPHOS.

Differences in the transcription level between nuclear and mitochondrial subunits were assessed by plotting the distribution of log₂ (FPKM), and the Wilcoxon rank-sum test was performed for statistical support. A hierarchical clustering analysis (Euclidean distance, Ward's method) was applied to generate transcription level (FPKM) heatmaps of both nuclear and mitochondrial subunits in each sample. Spearman's rank correlation coefficient was calculated across

transcription levels of males and females for each species and across males and females between species.

In order to detect sex- and/or species-specific co-transcriptional patterns of OXPHOS subunits, we calculated correlation matrices separately for the following conditions: *R. decussatus* males, *R. decussatus* females, *R. philippinarum* males, *R. philippinarum* females. More in details, for each pair of OXPHOS subunits, we calculated the Spearman's correlation coefficient to estimate the transcriptional correlation among all samples belonging to a given condition. Matrices were visualized through heatmaps.

Rates of protein evolution of OXPHOS subunits

We used MUSCLE (Edgar 2004) to align orthologous protein sequences, then protein alignments were back-translated into nucleotides using a custom R script. KaKs_Calculator 2.0 (Wang et al. 2010) was used to obtain the ratio of nonsynonymous to synonymous nucleotide substitution (dN/dS) between *R. decussatus* and *R. philippinarum*. Since males and females of *R. philippinarum* have different mitochondrial genomes, we reported two distinct dN/dS for mitochondrial subunits: one referred to male mitochondrial subunits (dN/dS between *R. decussatus* and *R. philippinarum* M-type) and one referred to female mitochondrial subunits (dN/dS between *R. decussatus* and *R. philippinarum* F-type). We plotted the distribution of dN/dS of nuclear and mitochondrial complexes and the Wilcoxon test was performed to evaluate significant differences. In order to investigate the presence of correlation between dN/dS and transcription levels, we plotted $\log_2(\text{FPKM})$ of nuclear subunits and their dN/dS for both *R. decussatus* and *R. philippinarum*. Concerning mitochondrial subunits, we plotted $\log_2(\text{FPKM})$ of *R. philippinarum* males against dN/dS between *R. decussatus* and *R. philippinarum* M-type; likewise we plotted $\log_2(\text{FPKM})$ of *R. philippinarum* females against dN/dS between *R. decussatus* and *R. philippinarum* F-type. Finally, we chose to plot $\log_2(\text{FPKM})$ of *R. decussatus* mitochondrial subunits against dN/dS between *R. decussatus* and *R. philippinarum* F-type, since F-type is considered the ancestral mitochondrial genome (Zouros 2013). The Spearman's rank correlation coefficient was calculated for each comparison..

Rates of protein evolution of nuclear genes involved in mitochondrial processes

In order to identify genes involved in mitochondrial processes, we used AmiGO2 (Balsa-Canto et al. 2016) to select loci annotated with the GO term "Mitochondrion" and every associated child terms. When these genes were recognized as orthologs between the two species, we obtained dN/dS following the same pipeline we used for OXPHOS subunits. When orthologous genes had a dN/dS higher than 0.2, we used REViGO (Supek et al. 2011) to highlight the GO

annotation in both the species, and networks of GO terms were visualized using the application DyNet (Goenawan et al. 2016) from the Cytoscape App Store (Shannon et al. 2003; Lotia et al. 2013). When genes were not recognized as orthologs, we used REViGO to obtain the GO annotations of these species-specific loci involved in mitochondrial processes and GO terms networks we visualized with DyNet.

Results

Differential transcription of OXPPOS genes

Combining the annotation obtained from BLASTP, HMMER, and Interpro2GO, we retrieved 40 nuclear-encoded subunits involved in OXPPOS. All mtDNA-encoded subunits were retrieved, but *atp8* was not included in the analyses because of known annotation/alignment issues of this gene in bivalves (Breton et al. 2010).

The transcription level of OXPPOS subunits in the two species is reported in Supplementary tables 1 and 2. The transcription level of nuclear-encoded OXPPOS subunits is more correlated between males and females within species (*R. decussatus* = 0.87, Supplementary figure 1, green dots; *R. philippinarum* = 0.92, Supplementary figure 1, blue dots), and less between males (Spearman's rank correlation = 0.66, Supplementary figure 1, orange dots) and females (Spearman's rank correlation = 0.5, Supplementary figure 1, red dots) between species. Compared with nuclear subunits, the transcription level of mtDNA-encoded subunits is less correlated between males and females within species (*R. decussatus* = 0.69, Supplementary figure 1, dark green triangles; *R. philippinarum* = 0.67, Figure 1, dark blue triangles), while it shows a higher correlation between males of the two species (Spearman's rank correlation = 0.67, figure 1, dark orange triangles), and between females (Spearman's rank correlation = 0.94, Supplementary figure 1, dark red triangles).

We compared the transcription levels of nuclear- and mtDNA-encoded orthologous genes involved in OXPPOS. In both *R. decussatus* and *R. philippinarum*, the transcription of mitochondrial subunits is markedly higher than that of nuclear subunits (figure 1). In *R. decussatus* the median transcription level of mitochondrial subunits is 19-fold higher than that of nuclear subunits (3,596 FPKM as opposed to 189 FPKM, respectively; Wilcoxon test p-value = 9.54E-07). In *R. philippinarum*, the median transcription level of mitochondrial subunits is 121-fold higher than that of nuclear subunits (5,101 FPKM vs 42 FPKM, respectively; Wilcoxon test p-value = 1.431E-06). Considering males and females separately, mitochondrial subunits are 12 times more transcribed than nuclear subunits in males of *R. decussatus* (Wilcoxon test p-value = 5.855E-05), 36 times more transcribed in females of *R. decussatus* (Wilcoxon test p-value = 6.042E-07), 110 times more transcribed in males of *R. philippinarum* (Wilcoxon test p-value = 1.431E-06), and 131 times more transcribed in females of *R. philippinarum* (Wilcoxon test p-value = 9.54E-07) (Supplementary figure 2). In nuclear complexes, the transcription level seems to be more variable between males and females of *R. decussatus*—particularly for Complex II, IV and ATPase—while it is more similar between in males and females of *R. philippinarum* (Supplementary figure 3A). On the contrary, the transcription level of mitochondrial complexes is similar among males of *R. philippinarum* and

males and females of *R. decussatus*, while females of *R. philippinarum* are characterized by lower and more variable log₂(FPKM) of Complex I and higher log₂(FPKM) of Complex IV (Supplementary figure 3B).

We performed a hierarchical clustering analysis of transcription level in nuclear subunits of males and females of *R. decussatus* and *R. philippinarum* (figure 2). While many of these subunits seem to have similar levels of transcription in both the species, some subunits of Complex I, IV and ATPase are much more transcribed in males of *R. decussatus*. So, in *R. philippinarum* the transcription level is more uniform between males and females and generally lower than that in *R. decussatus*, where males tend to have a higher transcription compared to females. This pattern results in five of six males of *R. decussatus* to cluster separately from all other samples. While the cluster analysis was able to separate both sexes and species based on their transcription level, it did not cluster together subunits belonging to the same complex.

The same hierarchical clustering analysis was performed on mitochondrial subunits (figure 3): in this case, females of *R. philippinarum* cluster separately from other samples, due to the higher transcription of subunits of Complex IV (*cox1*, *cox2*, *cox3*), Complex III (*cytb*) and one subunit of Complex I (*nad4*) and to the lower transcription of some subunits belonging to Complex I (*nad4L*, *nad2*, *nad5*, *nad3*). On the contrary, the transcription level of most subunits is comparable among males and females of *R. decussatus* and males of *R. philippinarum*, so that the cluster analysis could not always cluster separately the two species. Also in this case, subunits were not clustered together based on Complex they belong to.

When we considered all OXPHOS subunits in the cluster analysis, we found that mitochondrial subunits tend to cluster separately from almost all nuclear subunits, due to their higher transcription level (Supplementary figure 4). Among samples, females of *R. philippinarum* cluster separately from all other samples because of the high variability in their mitochondrial subunits transcription, then males of *R. decussatus* cluster separately from other samples due to their higher transcription level of some nuclear subunits.

We performed correlation heatmaps of OXPHOS subunits transcription (figure 4), separately for males and females of both species. Correlation coefficients among subunits are reported in Supplementary table 3.

Evolutionary rates of OXPHOS genes

The ratio of nonsynonymous to synonymous nucleotide substitution (dN/dS) was calculated between nuclear subunits of *R. decussatus* and *R. philippinarum*, between mitochondrial subunits of *R. decussatus* and *R. philippinarum* F-type, and between mitochondrial subunits of *R. decussatus* and *R. philippinarum* M-type. As shown in figure 5, dN/dS of nuclear and mitochondrial subunits is pretty uniform (median nuclear = 0.14; median *R. decussatus* vs *R. philippinarum* F-type = 0.11; median *R. decussatus* vs *R. philippinarum* M-type = 0.15), and the differences are not statistically significant. Among nuclear subunits (figure 6, Supplementary table 4), dN/dS is variable, with highest values in Complex III (median = 0.28) and lowest values in Complex II (median = 0.04). Among mitochondrial subunits (figure 7), Complexes IV and ATPase have the highest dN/dS in the *R. decussatus* vs *R. philippinarum* M-type comparison (dN/dS \geq 0.15; figure 7A), and *R. decussatus* vs *R. philippinarum* F-type comparison (figure 7B; dN/dS \geq 0.15), while Complex III is characterized by the lowest value (dN/dS=0.08; figure 7A,B) (Supplementary table 4).

Correlation between transcription level and evolutionary rate in OXPHOS genes

Figure 8 shows the relationship between log₂(FPKM) and dN/dS in nuclear and mitochondrial subunits. We did not find a negative correlation between transcription level and dN/dS of nuclear subunits neither in *R. decussatus* (Spearman's correlation = 0.13, p-value = 0.41; figure 8A, circles), nor in *R. philippinarum* (Spearman's correlation = 0.33, p-value = 0.03; figure 8B, circles). At the same way, we did not find a statistically supported correlation between transcription level and rate of protein evolution in mitochondrial subunits, neither in *R. decussatus* (Spearman's correlation = 0.26, p-value = 0.4; figure 8A, black triangles) nor in *R. philippinarum* M-type (Spearman's correlation = -0.16, p-value = 0.6; figure 8B, empty triangles), nor in *R. philippinarum* F-type (Spearman's correlation = 0.09, p-value = 0.7; figure 8B, black triangles).

Annotation and differential transcription of genes involved in mitochondrial biology

We found 1,264 loci in *R. decussatus* and 1,159 in *R. philippinarum* annotated either with the GO term 'Mitochondrion' or any of its child terms. Of these loci, 747 in *R. decussatus* and 696 in *R. philippinarum* were recognized as orthologs between the two species, while 148 were specific of *R. decussatus* and 141 were specific of *R. philippinarum*.

Among orthologous genes annotated with the GO term “Mitochondrion”, we retrieved 51 loci with a rate of protein evolution >0.2 . Based on the GO annotation, such genes are involved in regulation of transcription, translation, DNA recombination, respiratory chain assembly, and response to oxidative stress (figure 9). A list of the most abundant GO terms is also reported in Supplementary table 5. In particular, the two genes with the highest rate of protein evolution ($dN/dS = 0.5$) are involved in mitochondrial translation initiation, and apoptotic signaling.

Concerning species-specific genes, both in *R. decussatus* and *R. philippinarum*, GO annotation seems to be mainly related to mitochondrion organization, fission and fusion, and apoptotic changes (Figure 10). Several GO terms were found exclusively in *R. decussatus* (Figure 10, green nodes, Supplementary table 6) or *R. philippinarum* (Figure 10, red nodes, Supplementary table 6), and they are mainly involved in regulation of mitochondrion organization (e.g.: localization, inner membrane organization, mitochondrial fission, sperm mitochondrion organization), in calcium ion homeostasis, and in regulation of mitochondrial DNA replication, transcription and translation.

Discussion

In this work we investigated the rate of protein evolution and the transcription level of nuclear and mtDNA-encoded OXPHOS subunits in the bivalve species *Ruditapes philippinarum* and *Ruditapes decussatus*. *R. philippinarum* share with ~100 bivalve species a peculiar mitochondrial inheritance mechanism: two mitochondrial lineages are present in these animals, one—the F-type—is transmitted by mothers to all the progeny, and one—the M-type—is transmitted by fathers to males only. Therefore, *R. philippinarum* is naturally heteroplasmic and its nuclear genome had to co-evolve with two different mitochondrial genomes. On the contrary, *R. decussatus* is characterized by a standard maternal inheritance of mitochondria, where the paternal mitochondrial contribution is avoided. Therefore, the comparison of nuclear and mitochondrial subunits between these two species offers an exceptional opportunity to investigate the dynamics of mito-nuclear evolution, and the response of nuclear genome to a natural condition of high mitochondrial variability.

Rate of protein evolution of OXPHOS subunits

The rate of protein evolution (dN/dS) of mtDNA-encoded OXPHOS subunits was largely investigated across Metazoa (see for example Popadin et al. 2012; Piganeau & Eyre-Walker 2009; Nabholz et al. 2012), commonly revealing values very close to zero. To explain this strong purifying selection, it was hypothesized that mitochondrial subunits are subject to strict functional constraints, since they assemble the core of OXPHOS complexes, and the accumulation of nonsynonymous substitution may compromise the proper protein-protein recognition. Nevertheless, evidence of higher dN/dS of mitochondrial OXPHOS subunits was reported as well, either due to positive or relaxed selection acting on mitochondrial genes. On the one hand, Bazin et al. (2006) reported evidence of positive selection on mtDNA, particularly in invertebrates, where over 60% of nonsynonymous substitutions are likely fixed (James et al. 2016). On the other hand, other studies suggested that species with high energy needs due to their locomotive habits are characterized by a lower dN/dS of mitochondrial OXPHOS subunits, compared to species with lower energy needs. On the contrary, mitochondrial OXPHOS subunits of species with a more sedentary life have a higher dN/dS (Shen et al. 2009; Strohm et al. 2015; Mitterboeck & Adamowicz 2013; Chong & Mueller 2013). Considering the lower energy needs, the Authors hypothesized a relaxed purifying selection acting on mitochondrial OXPHOS subunits, rather than an adaptive evolution.

Here, we show that dN/dS of mitochondrial OXPHOS subunits between *R. decussatus* and *R. philippinarum* is an order of magnitude higher compared to that of most animal taxa investigated

so far (Nabholz et al. 2012; Popadin et al. 2012), both when we consider *R. philippinarum* F-type and *R. philippinarum* M-type in the comparison with *R. decussatus* (figure 5). Therefore, apparently mitochondrial subunits in these species are allowed to be more variable and nonsynonymous substitutions accumulate easier compared to most taxa investigated so far. Considering OXPHOS complexes separately, dN/dS ranges between 0.1 and 0.2, with higher values in Complex IV and ATPase e lower in Complex I and Complex III in both the *R. decussatus* vs *R. philippinarum* M-type comparison (figure 7A), and *R. decussatus* vs *R. philippinarum* F-type comparison (figure 7B). As mentioned before, higher dN/dS could reflect either positive or relaxed selection. On the one hand, here we investigated dN/dS of mtDNA-encoded subunits in two bivalve species that lead a very sedentary life, spending most of the time buried in the sand. Therefore, the higher dN/dS could be the result of a relaxation in functional constraints, due to the sedentary life of these species. On the other hand, it should be highlighted that the two species investigated are characterized by different mechanisms of mitochondrial inheritance and two mtDNA variants, which are very different in sequence, co-exist in *R. philippinarum* (amino acid p-distance = 34%). Also, it was assumed that mtDNA genomes of DUI species undergo more changes and at a faster rate than the genomes of species with standard maternal inheritance, and that the two mitochondrial variants could have sex-specific functions (Zouros 2013). Therefore, the higher dN/dS could be due to a positive selection acting on mtDNA in DUI species. Further analyses including more species—both DUI and SMI—could help to clarify this point. Alternatively, it has been recently proposed that transcription level is the main determinant in affecting the rate of protein evolution of mitochondrial subunits ((see for example: Nabholz et al. 2012; Popadin et al. 2012)). For a detailed discussion about this topic, see the paragraph ‘The relationship between transcription level and rate of protein evolution in OXPHOS subunits’.

Few works investigated the rate of protein evolution of both nuclear and mtDNA-encoded OXPHOS subunits and the coevolution between these two genomes ((see for example: Nabholz et al. 2012; Popadin et al. 2012); (Burton et al. 2006; Dowling et al. 2008; Osada & Akashi 2012; Havird & Sloan 2016; Aanen et al. 2014; Burton & Barreto 2012); (Nabholz et al. 2012; Popadin et al. 2012)). In particular, most of these studies focused on model-organisms such as primates, mice, and copepods, and investigated the effects of mitochondrial heteroplasmy by producing cytoplasmic hybrids. One of the main hypotheses about the dynamics of mito-nuclear coevolution is the ‘nuclear compensation hypothesis’, which posits that nuclear subunits evolve faster to compensate for the high mutation rate of mtDNA, and ensure the proper recognition among OXPHOS subunits. In this work we investigated the nuclear response to a natural condition of heteroplasmy and to the above discussed high dN/dS of mitochondrial subunits. According to the nuclear compensation hypothesis, we expected to see an increase of dN/dS of nuclear subunits in response to the higher rate of protein evolution of mitochondrial genes in the species we investigated. Surprisingly, while dN/dS of the mitochondrial subunits is an order of magnitude higher compared to species where mito-nuclear coevolution was investigated so far,

dN/dS of nuclear subunits is comparable to that of other species; furthermore, there are no statistically significant differences in the distribution of dN/dS between nuclear and mtDNA-encoded subunits (figure 5). Therefore, no evidence for nuclear compensation was found in these species in response to the higher rate of evolution of mitochondrial genome. In addition, this lack of compensation is remarkably evident when we investigate the evolution of sequences for each OXPHOS complex separately: among nuclear subunits, Complex III has the highest dN/dS (median dN/dS = 0.28), on the contrary Complex III has the lowest rate of protein evolution among mitochondrial complex (median dN/dS = 0.08); furthermore in Complex IV and ATPase the dN/dS is higher in mitochondrial subunits compared to nuclear subunits (figure 6 and 7, Supplementary table 4). Similar results were obtained by analyzing synonymous substitution rate (dS) and nonsynonymous substitution rate (dN) of OXPHOS and non-OXPHOS genes in vertebrates (Zhang & Broughton 2013). The authors found that dN of mitochondrial subunits is not always higher than that of nuclear subunits in each complex, and suggested a minor role for compensatory mechanism in the evolution of OXPHOS genes. Although many works support the importance of a proper interaction among OXPHOS subunits, the present study rather suggests an independent trend of nuclear and mitochondrial OXPHOS complexes concerning rates of protein evolution. In any case, it should be considered that these proteins are under purifying selection, with values of dN/dS that are mostly below 0.3. Also, while many works report dN/dS in OXPHOS subunits, few works have examined the effects of these mutations on proteins function (da Fonseca et al. 2008; Azevedo et al. 2009; Schmidt et al. 2001). Therefore, it is possible that OXPHOS proteins can tolerate mutations whilst maintaining their function intact (i.e.: robustness, see Kitano 2004), particularly if such mutations affect domains not directly involved in the interactions among subunits, or if such mutations are compensated by other nonsynonymous substitutions. In addition, even a nonsynonymous substitution could have no effects on the protein structure thus yielding a working complex, particularly in species with low metabolic requirements. Nevertheless, mitochondrial genome of the species we investigate not only has relatively high dN/dS but also more variants exist within *R. philippinarum* (F type and M type). Many works highlight the negative effects of having different mtDNA variants within the same organism, and uniparental inheritance of mitochondria is thought to have evolved in order to avoid mixing different mitochondrial lineages (see for example Lane 2012). So far, most of the works about mito-nuclear incompatibilities investigated the effects of heteroplasmy by creating cytoplasmic hybrids and the consequent reduction or breakdown of OXPHOS was ascribed to a mismatch between nuclear and mitochondrial subunits. Here we see that, in a natural condition of heteroplasmy, nuclear-encoded OXPHOS subunits do not seem to be affected by the mitochondrial variability that characterized these species, at least not at the DNA level. Still, other nuclear genes could be likely involved in mito-nuclear incompatibilities, like those arising in heteroplasmic hybrids, as proposed by Ellison and Burton (2008). It is known that there are ~1,500 nuclear genes involved in mitochondrial biology (Wallace 2005), that have to constantly interact with mitochondrial

genome, RNAs, and proteins, and some of these genes could be responsible for mito-nuclear incompatibilities. Since the nuclear genome of *R. decussatus* have to interact with only one mitochondrial variant, while the nuclear genome of *R. philippinarum* evolved interacting with two different mitochondrial lineages, these species provide a unique chance to investigate what genes, if any, evolved faster in response to mitochondrial variability. For this purpose, we investigated orthologous nuclear genes with putative mitochondrial target and dN/dS higher than 0.2; also, we detected species-specific loci involved in mitochondrial functions. We found genes involved in regulation of transcription, translation, DNA recombination, respiratory chain assembly, and response to oxidative stress (figure 9). In particular, we found two genes with dN/dS of 0.5 involved in mitochondrial translation initiation, and apoptotic signal. GO terms involved in mitochondrial regulation of transcription and translation were also found among species-specific nuclear genes (figure 10). In addition, GO terms involved in mitochondrial organization—including mitochondrial fission and fusion, fragmentation, mitochondrial localization, and apoptosis—were largely represented in specie-specific loci. Ellison and Burton (2008) already proposed that nuclear genes involved in mitochondrial transcription could be responsible for mito-nuclear incompatibilities, and thus mainly involved in mito-nuclear coevolution. Here we report that genes involved in regulation of transcription and translation, as well as respiratory chain assembly factors, have a high rate of protein evolution in these species and could be subject to a faster evolution in response to the exceptional mtDNA variability.

Transcription level of mitochondrial and nuclear OXPHOS subunits

We found that transcription level of mtDNA-encoded OXPHOS subunits is much higher than that of nuclear-encoded subunits in both the species. This pattern was already reported in a wide range of eukaryotes, where the transcript abundance of mitochondrial subunits is 20-fold higher than that of nuclear subunits in animals, 18-fold higher in plants and 6-fold higher in fungi (Nabholz et al. 2012; Havird & Sloan 2016). Here, we report a high abundance of mitochondrial transcripts in *R. decussatus*, and a particularly high abundance in *R. philippinarum*, that is, respectively, 19-fold and 121-fold higher than the transcription of nuclear subunits (figure 1). The reason of this remarkable difference in transcript abundance between mitochondrial and nuclear subunits it is not clear. This pattern was proposed to be ascribed to peculiar property of mitochondrial transcription machinery, and different hypotheses were taken in consideration, such as inefficient mitochondrial translation (Woodson and Chory, 2008; Havird and Sloan, 2016).

It is not clear if the transcription level is correlated among OXPHOS subunits; previous works found patterns of co-transcription among subunits belonging to the same OXPHOS complex (van Waveren & Moraes 2008; Garbian et al. 2010). We found that the cluster analysis can usually separate samples based on species and sex. In particular, when we consider nuclear subunits, males of *R. decussatus* cluster apart for their higher transcription in some subunits of Complex I,

Complex IV, and ATPase (figure 2). A different pattern is shown when we consider mitochondrial subunits, where females of *R. philippinarum* cluster separately from all the other samples due to their higher transcription level, then females of *R. decussatus* cluster separately from males of the two species, which cluster together (figure 3). Interestingly, the cluster analysis is not able to separate subunits based on their complex affiliation in any case. This observation suggests that transcript abundance is considerably variable among subunits, and no pattern of complex-specific co-transcription is found in these species. In addition, we performed correlation matrices among subunits, separately for males and females in each species. Even in this case, a lack of complex-specific correlation pattern is evident in each condition (figure 4, Supplementary figures 5-8). This is interesting, because most of the OXPHOS subunits exist at same ratio in all complexes, except for some subunits of the ATPase (Hüttemann et al. 2007). Concerning mitochondrial subunits, it should be highlighted that the knowledge about mitochondrial transcription, as well as post-transcriptional regulatory mechanisms and protein turnover, is very restricted (Sirey & Ponting 2016). Most of the studies about these topics focus on mammals (Asin-Cayuela & Gustafsson 2007) and, in any case, many genes and mechanisms remain uncharacterized. Even more surprisingly, we still do not know the role of polyadenylation in mitochondrial transcripts. It is commonly known that poly(A) tail confers stability to cytosol transcripts, ensures their exit from the nucleus and allows the initiation of translation (Rorbach & Minczuk 2012). On the contrary, a poly(A) tail is required for transcript degradation in bacteria and plant mitochondria (Gagliardi et al. 2004). So, what is the role of poly(A) tail in animal mitochondrial transcripts? Does it confer stability, or is it rather required to degradation, similarly to prokaryotes? Slomovic et al. (2008) proposed that these two mechanisms could co-exist in mitochondria. Considering that the libraries for RNA-Seq experiments are constructed by isolating polyadenylated transcripts, knowing the role of the poly(A) tail is fundamental. On the one hand, if poly(A) tail is required for degradation, by estimating the abundance of polyadenylated transcripts we would count transcripts destined to degradation. On the other hand, if poly(A) tail is required for both stabilization and degradation, a quantification based on polyadenylated transcripts would represent an overestimation of the transcripts that will be eventually expressed. The convoluted dynamics of protein expression in mitochondria could then explain the absence of the expected correlation within OXPHOS complexes; if this is the case, we should also take into consideration the possibility that mitochondrial transcriptomics invariably yields noisy data.

That said, we observed a lack of co-transcription also among nuclear subunits. Even in nuclear genes, there is evidence of a low correlation between transcription level and protein abundance, and it was estimated that ~60% of the variation in protein concentration is due to post-transcriptional regulation (Vogel & Marcotte 2012).

The relationship between transcription level and rate of protein evolution in OXPHOS subunits

According to a rapidly spreading theory, transcription level is the main factor determining the rate of protein evolution, and highly-transcribed genes are characterized by lower dN/dS (Zhang & Yang 2015). Therefore there is a negative correlation (defined E-R correlation) between transcript abundance and dN/dS. Recently, the E-R correlation was taken in consideration to explain the low rate of protein evolution affecting mitochondrial OXPHOS subunits. Nabholz et al. (2012) found a negative correlation between transcription level of OXPHOS subunits and dN/dS in different organisms, and the Authors proposed that the high transcription level of mitochondrial genes is the main responsible for the low rate of protein evolution. In the present work we investigated the relationship between transcripts abundance and dN/dS in two bivalve species. This is an interesting analysis, since *R. philippinarum* has a particularly high transcription of mitochondrial subunits, that is 210-fold higher compared to nuclear subunits and 10-fold higher than that in other animal mitochondrial subunits. Therefore, according to E-R correlation, a remarkable low dN/dS should characterize mitochondrial subunits of *R. philippinarum*. Still, dN/dS of these subunits is high as well, compared to other animals, and it is instead comparable to *R. decussatus*, where the transcription of mitochondrial genes is considerably lower. In addition, we did not find any negative correlation, neither in mitochondrial nor nuclear subunits in both the species (figure 8). Similar results were obtained by Havird and Sloan (2016) in plants, and the Authors concluded that transcript abundance cannot be responsible for dN/dS. In Chapter 3 we already discussed about the need to be careful in searching a correlation between these two variables. Here we confirm no evidence for E-R correlation. Also, our results show that transcription level of both nuclear and mitochondrial subunits is highly variable, and it is not uniform even among samples within species. Our opinion is that dN/dS is not affected by transcripts abundance and we suggest that the role of transcription level in determining the rate of protein evolution should be reconsidered.

Funding

This work was supported by the Italian Ministry of Education, University and Research MIUR FIR2013 Programme [RBF13T97A funded to FG]; and by the Canziani bequest funded to MP.

References

- Aanen DK, Spelbrink JN, Beekman M. 2014. What cost mitochondria? The maintenance of functional mitochondrial DNA within and across generations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369. doi: 10.1098/rstb.2013.0438.
- Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 201500012.
- Asin-Cayuela J, Gustafsson CM. 2007. Mitochondrial transcription and its regulation in mammalian cells. *Trends Biochem. Sci.* 32:111–117.
- Azevedo L et al. 2009. Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. *BMC Genomics.* 10:266.
- Ballard JWO, Melvin RG, Katewa SD, Maas K. 2007. Mitochondrial DNA variation is associated with measurable differences in life-history traits and mitochondrial metabolism in *Drosophila simulans*. *Evolution.* 61:1735–1747.
- Balsa-Canto E, Henriques D, Gábor A, Banga JR. 2016. AMIGO2, a toolbox for dynamic modeling, optimization and control in systems biology. *Bioinformatics.* 32:3357–3359.
- Barreto FSS, Burton RSS. 2013. Elevated oxidative damage is correlated with reduced fitness in interpopulation hybrids of a marine copepod. *Proceedings of the Royal Society B: Biological Sciences.* 280:20131521.
- 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.
- Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. *Science.* 312:570–572.
- Birky CW Jr. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences.* 92:11331–11338.
- 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.
- Burton RS, Barreto FS. 2012. A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? *Mol. Ecol.* 21:4942–4957.
- Burton RS, Ellison CK, Harrison JS. 2006. The sorry state of F2 hybrids: consequences of rapid

- mitochondrial DNA evolution in allopatric populations. *Am. Nat.* 168:S14–S24.
- Camacho C et al. 2009. BLAST+: architecture and applications. *BMC Bioinformatics.* 10:421.
- Castellana S, Vicario S, Saccone C. 2011. Evolutionary patterns of the mitochondrial genome in Metazoa: exploring the role of mutation and selection in mitochondrial protein coding genes. *Genome Biol. Evol.* doi: 10.1093/gbe/evr040.
- Chong RA, Mueller RL. 2013. Low metabolic rates in salamanders are correlated with weak selective constraints on mitochondrial genes. *Evolution.* 67(3):894-9.
- Dowling DK, Friberg U, Lindell J. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends Ecol. Evol.* 23:546–554.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Ellison CK, Burton RS. 2008. Genotype-dependent variation of mitochondrial transcriptional profiles in interpopulation hybrids. *Proceedings of the National Academy of Sciences.* 105:15831.
- da Fonseca RR, Johnson WE, O'Brien SJ, Ramos MJ, Antunes A. 2008. The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics.* 9:119.
- Gagliardi D, Stepien PP, Temperley RJ, Lightowlers RN, Chrzanowska-Lightowlers ZM. 2004. Messenger RNA stability in mitochondria: different means to an end. *Trends Genet.* 20:260–267.
- Garbian Y, Ovadia O, Dadon S, Mishmar D. 2010. Gene expression patterns of oxidative phosphorylation complex I subunits are organized in clusters. *PLoS One.* 5:e9985.
- Gershoni M, Templeton AR, Mishmar D. 2009. Mitochondrial bioenergetics as a major motive force of speciation. *Bioessays.* 31:642–650.
- Ghiselli F et al. 2012. De Novo assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. *Mol. Biol. Evol.* 29:771–786.
- Ghiselli F et al. 2013. Structure, transcription, and variability of metazoan mitochondrial genome: perspectives from an unusual mitochondrial inheritance system. *Genome Biol. Evol.* 5:1535–1554.
- Ghiselli F et al. 2017. The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae). *PeerJ.* 5:e3692.
- 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.

Gibson JD, Niehuis O, Verrelli BC, Gadau J. 2010. Contrasting patterns of selective constraints in nuclear-encoded genes of the oxidative phosphorylation pathway in holometabolous insects and their possible role in hybrid breakdown in *Nasonia*. *Heredity* . 104:310–317.

Goenawan IH, Bryan K, Lynn DJ. 2016. DyNet: visualization and analysis of dynamic molecular interaction networks. *Bioinformatics*. 32:2713–2715.

Guerra D, Ghiselli F, Milani L, Breton S, Passamonti M. 2016. Early replication dynamics of sex-linked mitochondrial DNAs in the doubly uniparental inheritance species *Ruditapes philippinarum* (*Bivalvia Veneridae*). *Heredity* . 116:324–332.

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.

Havird JC, Sloan DB. 2016. The roles of mutation, selection, and expression in determining relative rates of evolution in mitochondrial vs. nuclear genomes. *Mol. Biol. Evol.* doi: 10.1093/molbev/msw185.

Hill GE. 2016. Mitonuclear coevolution as the genesis of speciation and the mitochondrial DNA barcode gap. *Ecol. Evol.* doi: 10.1002/ece3.2338.

Hill GE. 2017. The mitonuclear compatibility species concept. *Auk*. 134:393–409.

Hüttemann M, Lee I, Samavati L, Yu H, Doan JW. 2007. Regulation of mitochondrial oxidative phosphorylation through cell signaling. *Biochim. Biophys. Acta*. 1773:1701–1720.

James JE, Piganeau G, Eyre-Walker A. 2016. The rate of adaptive evolution in animal mitochondria. *Mol. Ecol.* 25:67–78.

Jones P et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 30:1236–1240.

Kazuno A-A et al. 2006. Identification of mitochondrial DNA polymorphisms that alter mitochondrial matrix pH and intracellular calcium dynamics. *PLoS Genet*. 2:e128.

Kitano H. 2004. Biological robustness. *Nat. Rev. Genet.* 5:826–837.

Lane N. 2012. The problem with mixing mitochondria. *Cell*. 151:246–248.

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

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–1760.

Lotia S, Montojo J, Dong Y, Bader GD, Pico AR. 2013. Cytoscape app store. *Bioinformatics*.

29:1350–1351.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science*. 311:1727–1730.

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.

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, Ghiselli F, Iannello M, Passamonti M. 2014. Evidence for somatic transcription of male-transmitted mitochondrial genome in the DUI species *Ruditapes philippinarum* (Bivalvia: Veneridae). *Curr. Genet.* 60:163–173.

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.* doi: 10.1093/nar/gkt263.

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

Moreno-Loshuertos R et al. 2006. Differences in reactive oxygen species production explain the phenotypes associated with common mouse mitochondrial DNA variants. *Nat. Genet.* 38:1261–1268.

Nabholz B, Ellegren H, Wolf JBW. 2012. High Levels of Gene Expression Explain the Strong Evolutionary Constraint of Mitochondrial Protein-Coding Genes. *Mol. Biol. Evol.* doi: 10.1093/molbev/mss238.

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.

Obata M, Sano N, Komaru A. 2011. Different transcriptional ratios of male and female transmitted mitochondrial DNA and tissue-specific expression patterns in the blue mussel, *Mytilus galloprovincialis*. *Dev. Growth Differ.* 53:878–886.

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.

Pavlova A et al. 2017. Purifying selection and genetic drift shaped Pleistocene evolution of the

- mitochondrial genome in an endangered Australian freshwater fish. *Heredity* . 118:466–476.
- Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS One*. 4:e4396.
- Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. 2012. Purifying Selection in Mammalian Mitochondrial Protein-Coding Genes Is Highly Effective and Congruent with Evolution of Nuclear Genes. *Mol. Biol. Evol.* doi: 10.1093/molbev/mss219.
- Rand DM, Haney RA, Fry AJ. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends Ecol. Evol.* 19:645–653.
- Rorbach J, Minczuk M. 2012. The post-transcriptional life of mammalian mitochondrial RNA. *Biochem. J.* 444:357–373.
- Sackton TB, Haney RA, Rand DM. 2003. Cytonuclear coadaptation in *Drosophila*: disruption of cytochrome c oxidase activity in backcross genotypes. *Evolution*. 57:2315–2325.
- Sato M, Sato K. 2013. Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. *Biochim. Biophys. Acta.* 1833:1979–1984.
- Schmidt TR, Wu W, Goodman M, Grossman LI. 2001. Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Mol. Biol. Evol.* 18:563–569.
- Shannon P et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:2498–2504.
- Sharpley MS et al. 2012. Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell*. 151:333–343.
- Shen Y-Y, Shi P, Sun Y-B, Zhang Y-P. 2009. Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. *Genome Res.* 19:1760–1765.
- Sirey TM, Ponting CP. 2016. Insights into the post-transcriptional regulation of the mitochondrial electron transport chain. *Biochem. Soc. Trans.* 44:1491–1498.
- Skibinski DO, Gallagher C, Beynon CM. 1994a. Mitochondrial DNA inheritance. *Nature*. 368:817–818.
- Skibinski DO, Gallagher C, Beynon CM. 1994b. Sex-limited mitochondrial DNA transmission in the marine mussel *Mytilus edulis*. *Genetics*. 138:801–809.
- Slomovic S, Portnoy V, Schuster G. 2008. Chapter 24 Detection and Characterization of Polyadenylated RNA in Eukarya, Bacteria, Archaea, and Organelles. In: *Methods in Enzymology*. Vol. Volume 447 Academic Press: Department of Biology Technion, Israel

Institute of Technology, Haifa, Israel. pp. 501–520.

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

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*. 6:e21800.

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

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13:227–232.

Wallace DC. 2005. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* 39:359–407.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genomics Proteomics Bioinformatics*. 8:77–80.

van Waveren C, Moraes CT. 2008. Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. *BMC Genomics*. 9:18.

Wolff JN, Ladoukakis ED, Enr'iquez JA, Dowling DK. 2014. Mitonuclear interactions: evolutionary consequences over multiple biological scales. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369. doi: 10.1098/rstb.2013.0443.

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 J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* 16:409–420.

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

Zouros E, Ball AO, Saavedra C, Freeman KR. 1994. Mitochondrial DNA inheritance. *Nature*. 368:818.

Zouros E, Oberhauser Ball A, Saavedra C, Freeman KR. 1994. An unusual type of mitochondrial DNA inheritance in the blue mussel *Mytilus*. *Proc. Natl. Acad. Sci. U. S. A.* 91:7463–7467.

Main figures

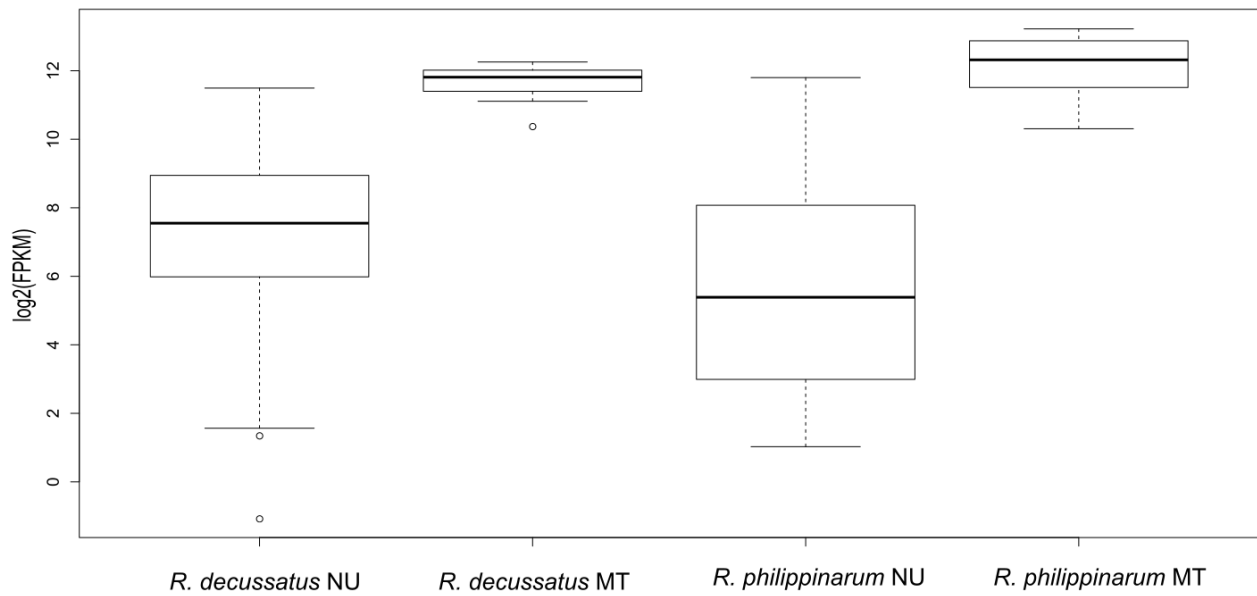


Figure 1: Transcription level of nuclear (NU) and mitochondrial (MT) OXPHOS subunits in *R. decussatus* and *R. philippinarum*.

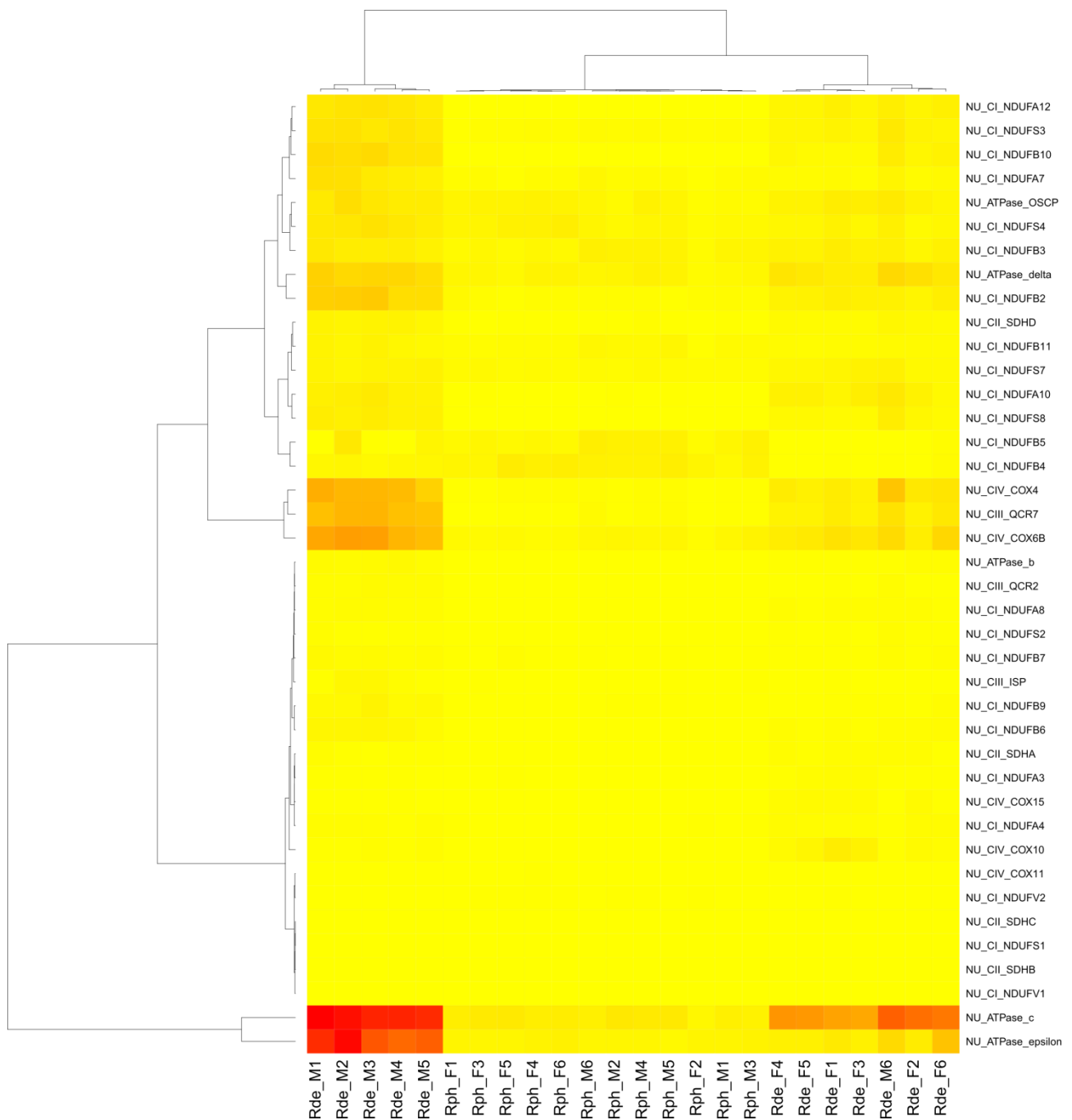


Figure 2: hierarchical clustering heatmap of nuclear OXPHOS subunits transcription in male (M1-M6) and female (F1-F6) samples of *R. decussatus* (Rde) and *R. philippinarum* (Rph).

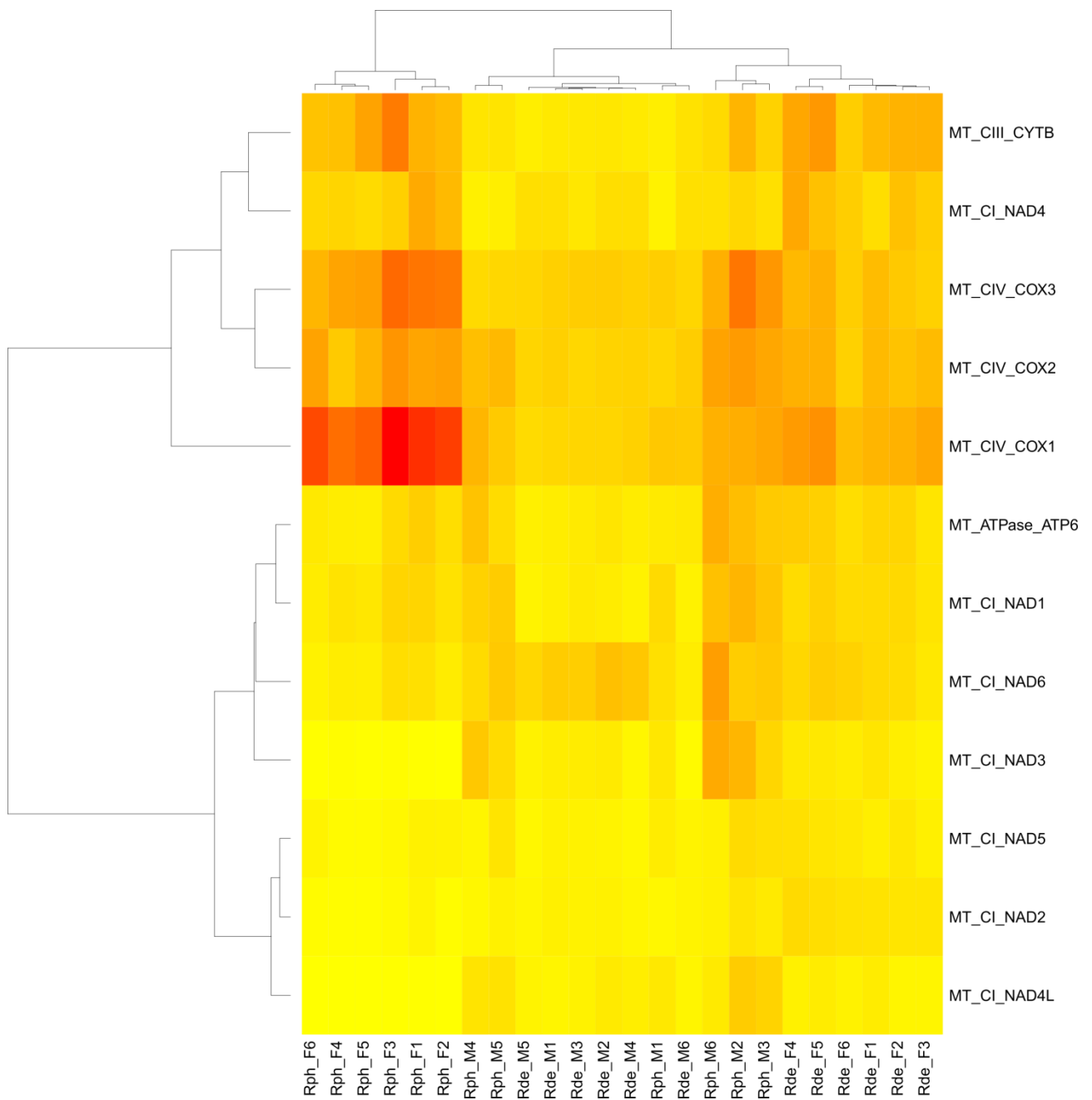


Figure 3: hierarchical clustering heatmap of mitochondrial OXPHOS subunits transcription in male (M1-M6) and female (F1-F6) samples of *R. decussatus* (Rde) and *R. philippinarum* (Rph).

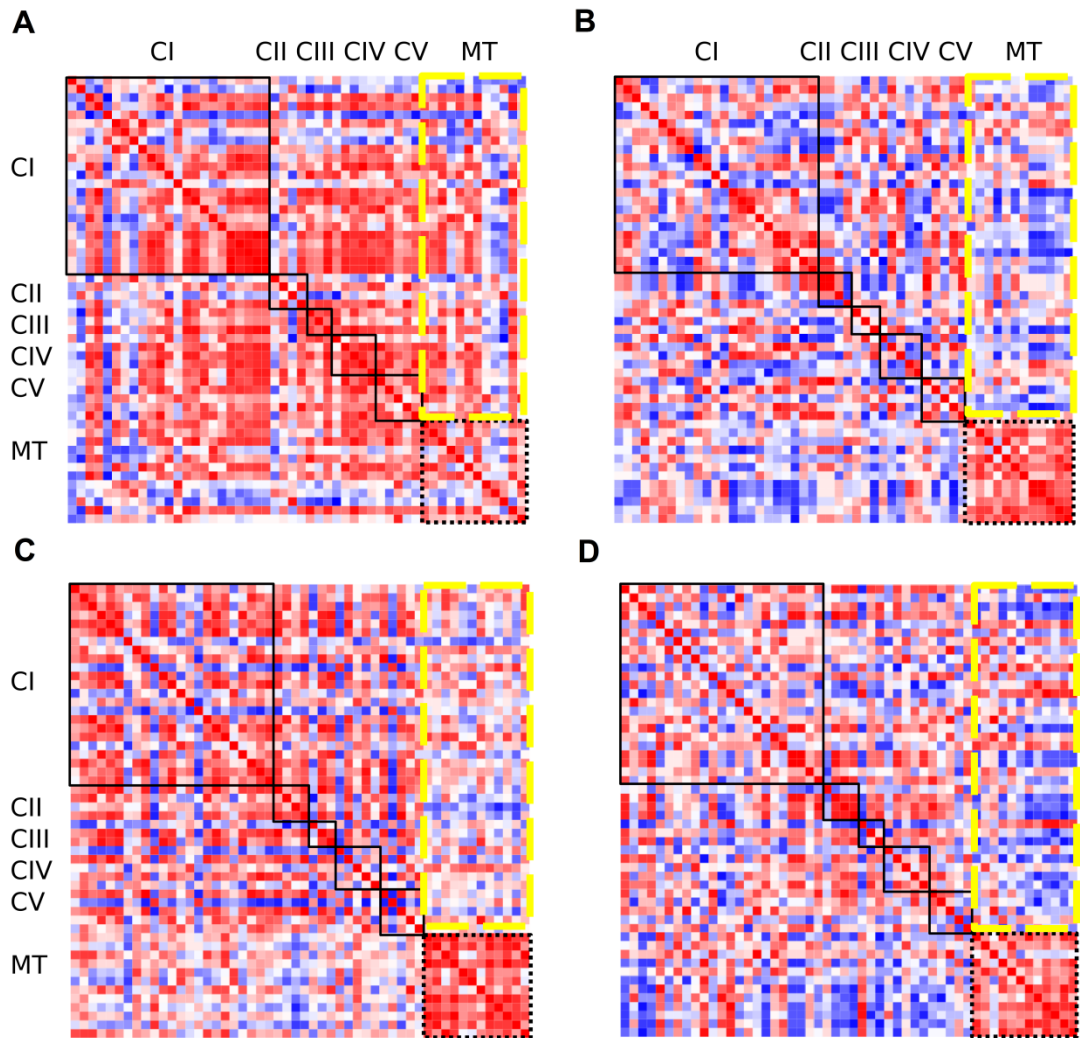


Figure 4: heatmaps of transcription correlation matrices among OXPHOS subunits in males and females of *R. decussatus* (respectively A and B) and males and females of *R. philippinarum* (respectively C and D). Red=positive correlation; Blue=negative correlation. Black solid squares=correlation among nuclear subunits within each OXPHOS complex; black dotted squares=correlation among mitochondrial subunits; yellow dashed squares=correlation among nuclear and mitochondrial subunits.

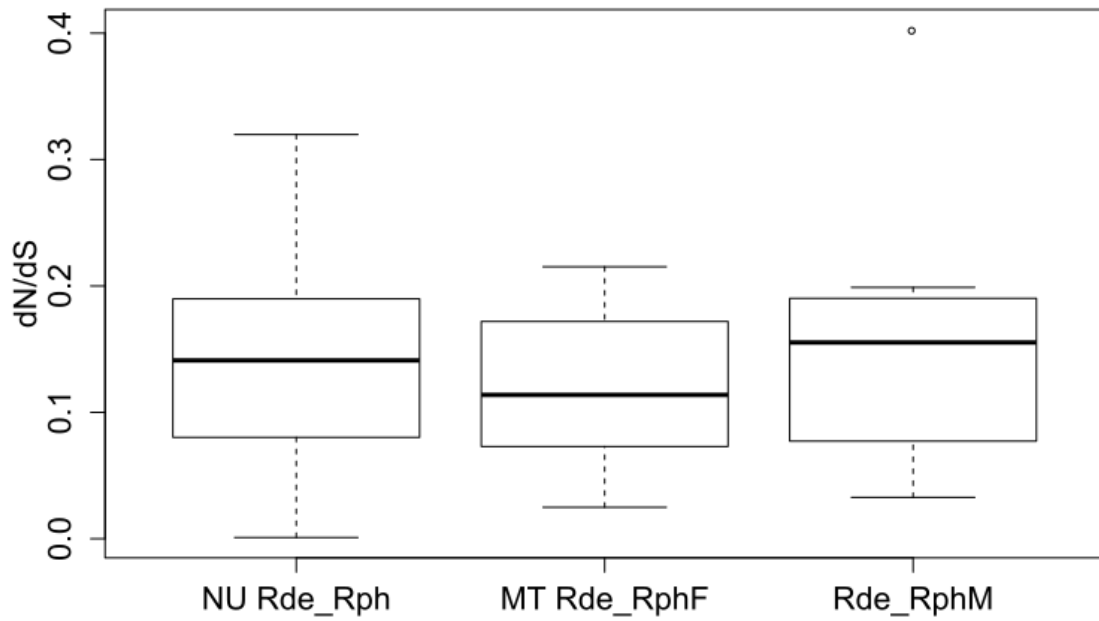


Figure 5: rate of protein evolution between nuclear OXPHOS subunits (NU) of *R. decussatus* and *R. philippinarum* (Rde_Rph) and between mitochondrial OXPHOS subunits (MT) of *R. decussatus* and *R. philippinarum* F-type (Rde_RphF) and *R. decussatus* and *R. philippinarum* M-type (Rde_RphM).

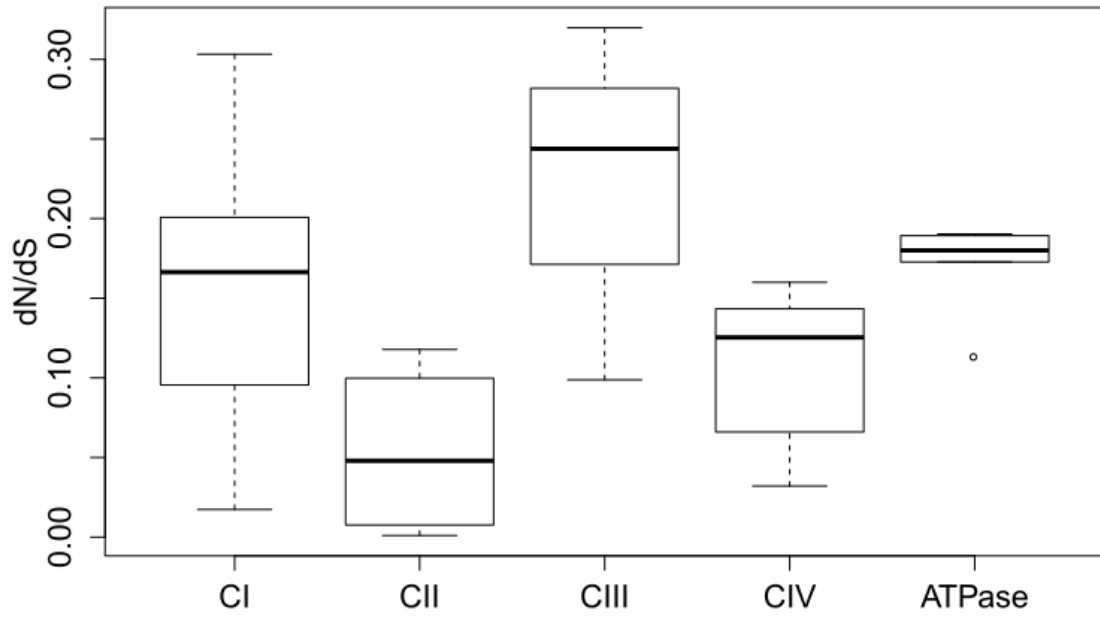


Figure 6: rate of protein evolution of nuclear subunits separately for each OXPHOS complex.

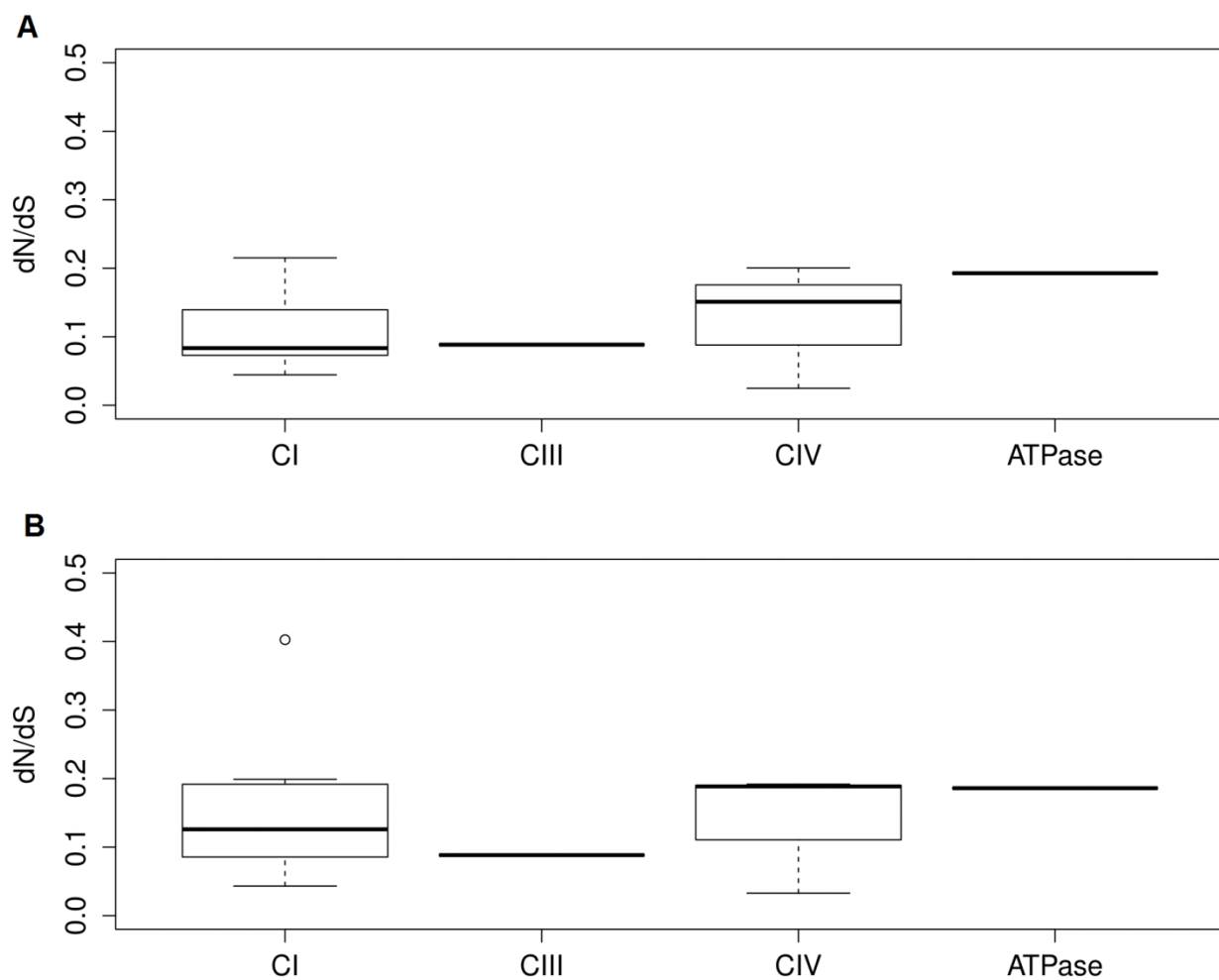


Figure 7: rate of protein evolution between mitochondrial OXPHOS subunits of *R. decussatus* and *R. philippinarum* M-type (A) and *R. decussatus* and *R. philippinarum* F-type (B) separately for each OXPHOS complex.

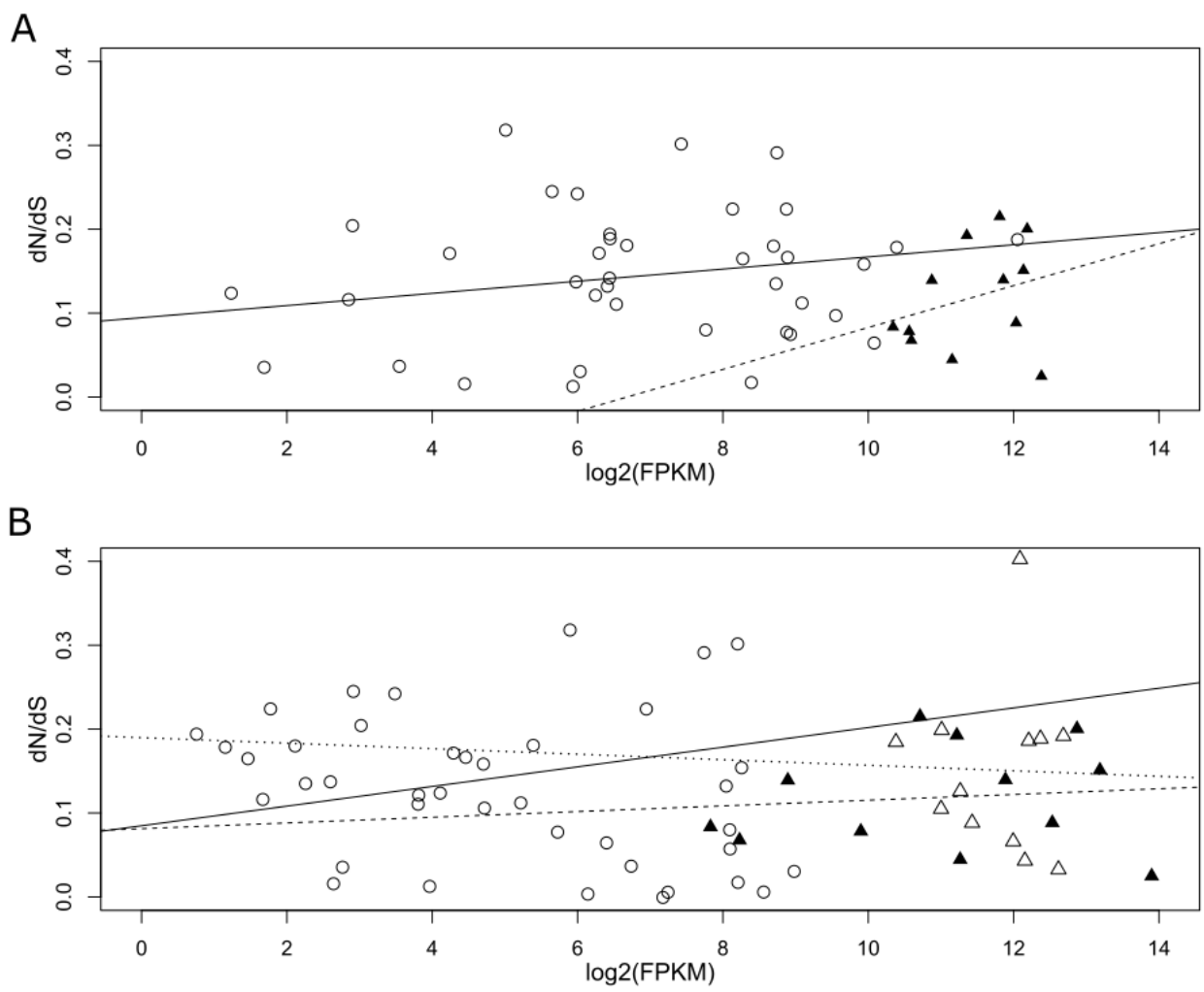


Figure 8: Relationship between transcription level ($\log_2(\text{FPKM})$) and rate of protein evolution (dN/dS) of OXPHOS subunits in *R. decussatus* (A) and *R. philippinarum* (B). Circles=nuclear subunits; black triangles= F-type mitochondrial subunits; empty triangles= M-type mitochondrial subunits. Lines represent linear regression among nuclear subunits (solid line), among F-type mitochondrial subunits (dashed line) and among M-type mitochondrial subunits (dotted line).

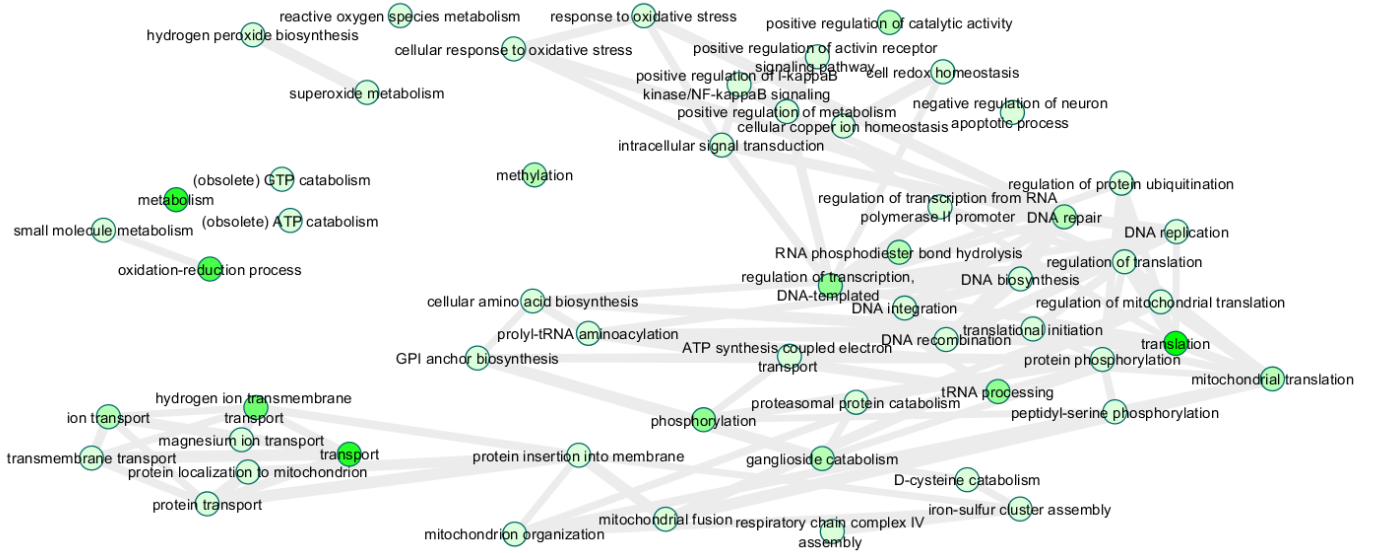


Figure 9: mitochondrial GO annotations of faster evolving genes ($dN/dS > 0.2$). Brighter colors indicate more represented GO terms; similar GO terms are linked by edges.

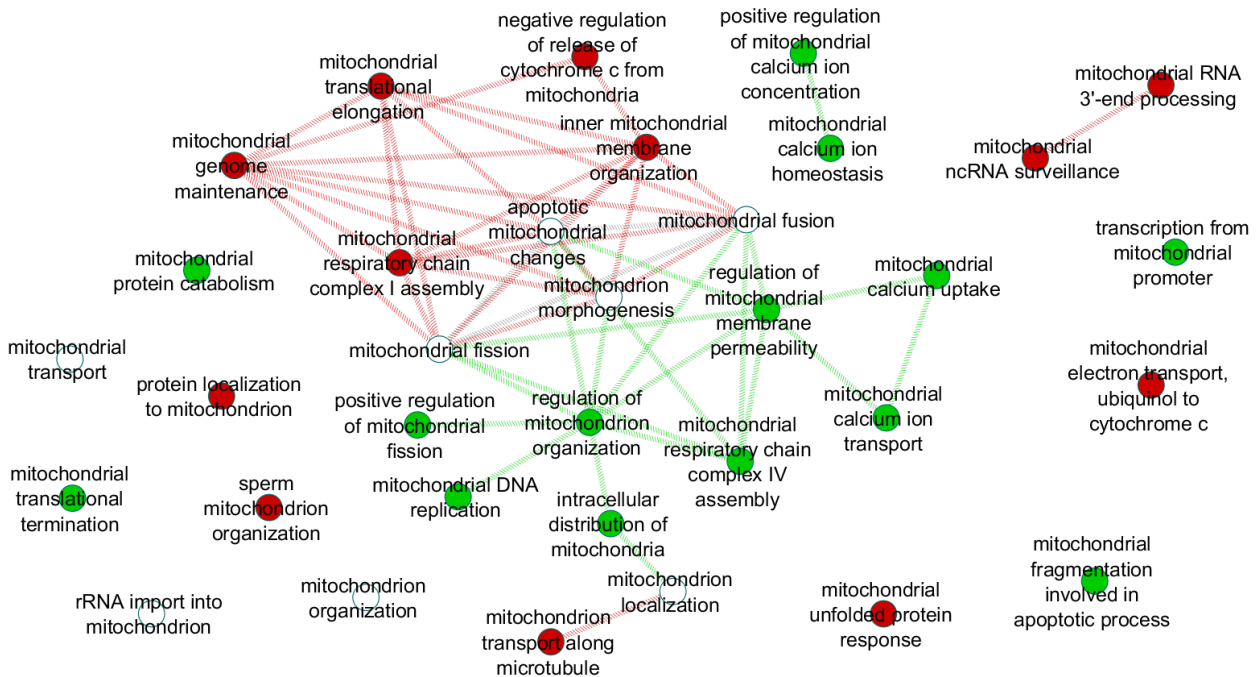
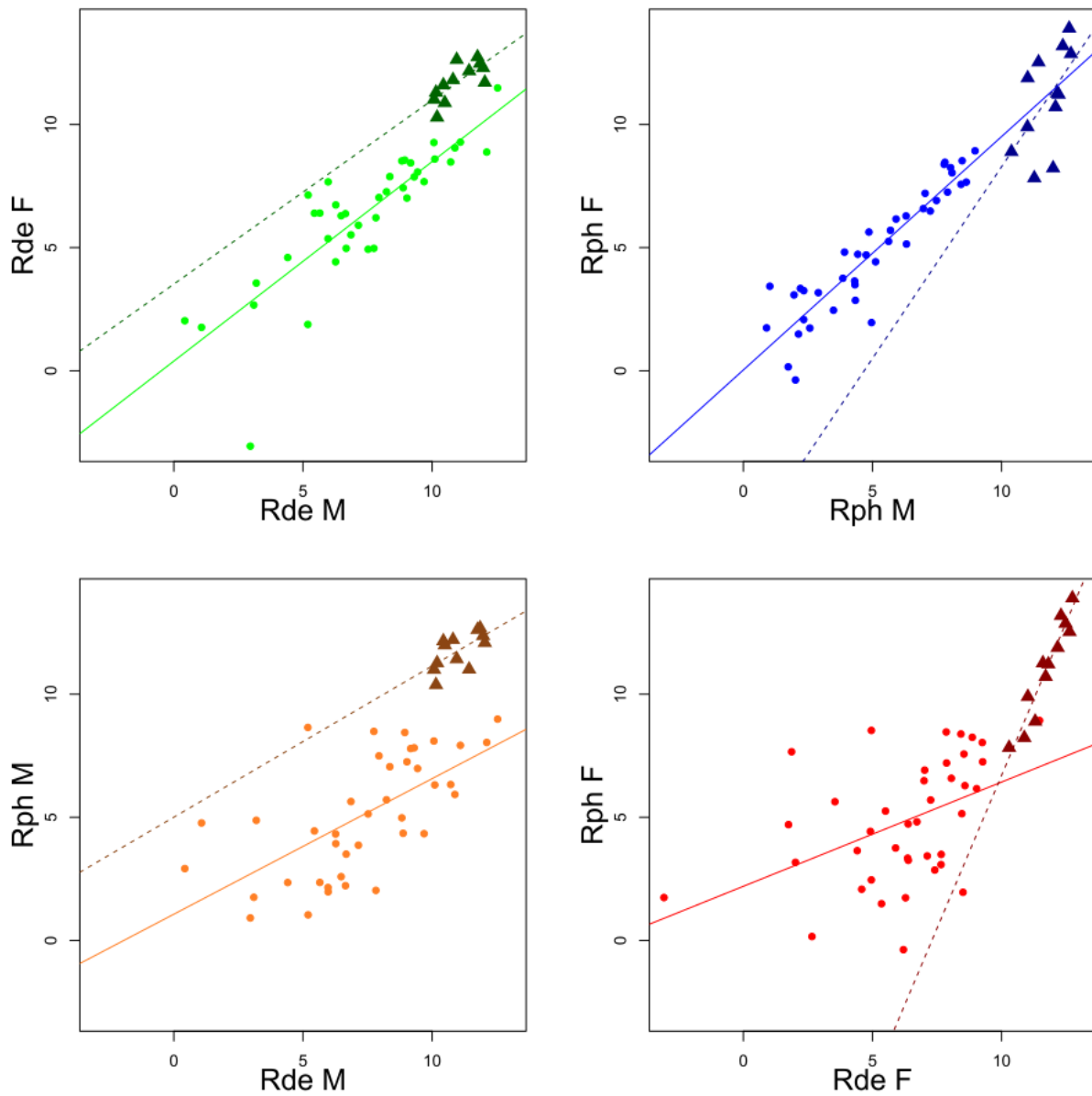
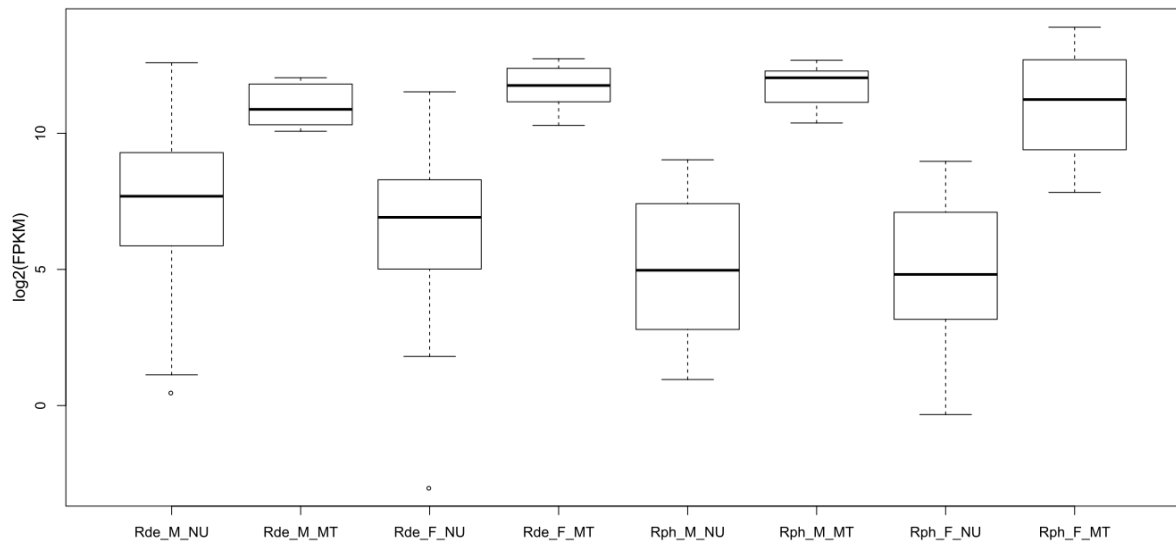


Figure 10: Mitochondrial GO annotations of *R. decussatus* (green) and *R. philippinarum* (red) specific genes. Shared annotations among species specific genes are visualized in white. Similar GO terms are connected by green edges in *R. decussatus* and red edges in *R. philippinarum*.

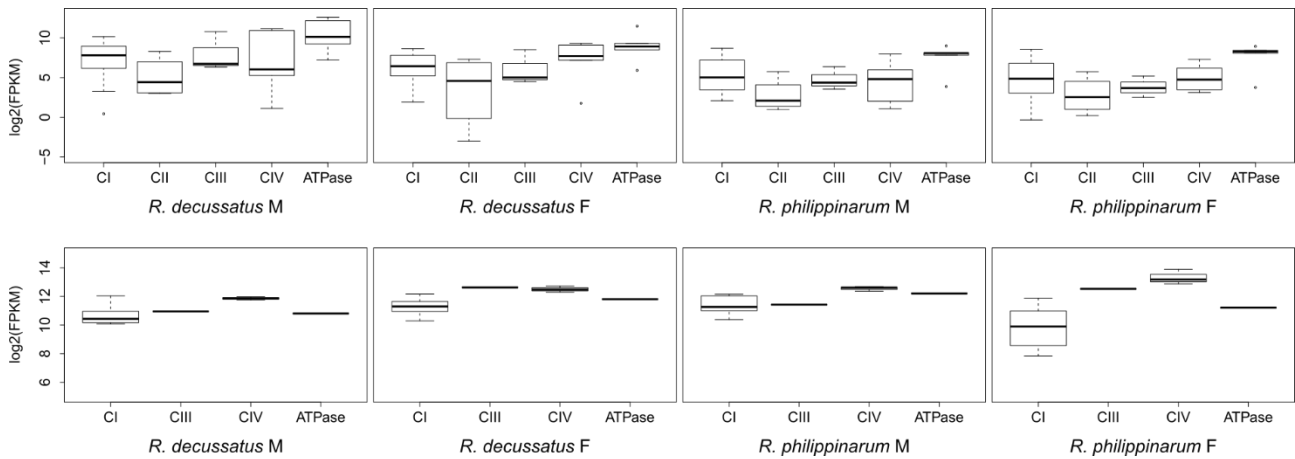
Supplementary materials



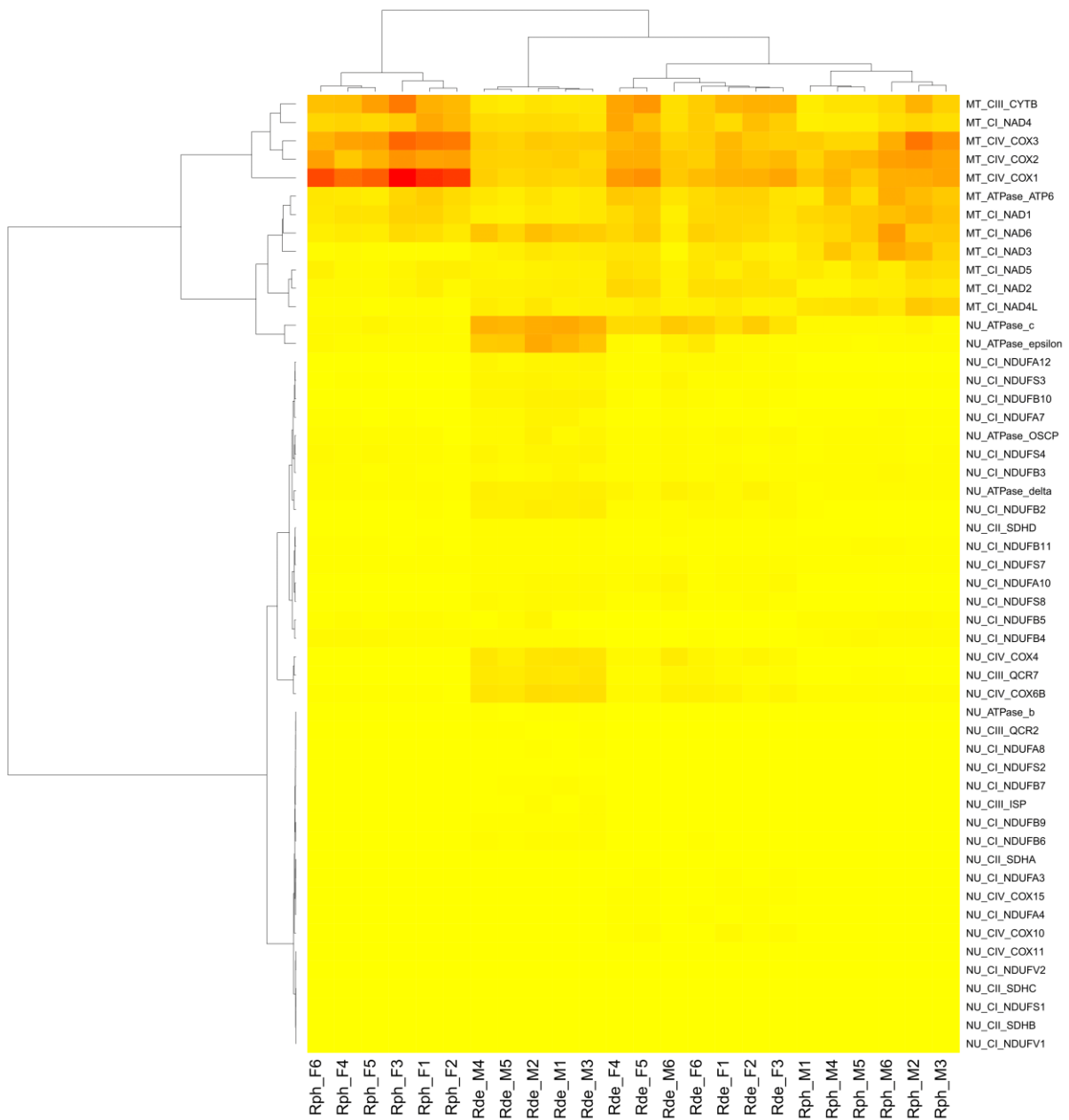
Supplementary figure 1: Correlation of OXPHOS subunits transcription levels between males and females within species (*R. decussatus*=green; *R. philippinarum*=blue), between males of *R. decussatus* and *R. philippinarum* (orange) and between females of *R. decussatus* and *R. philippinarum* (red). Dots=nuclear subunits; triangles=mitochondrial subunits.



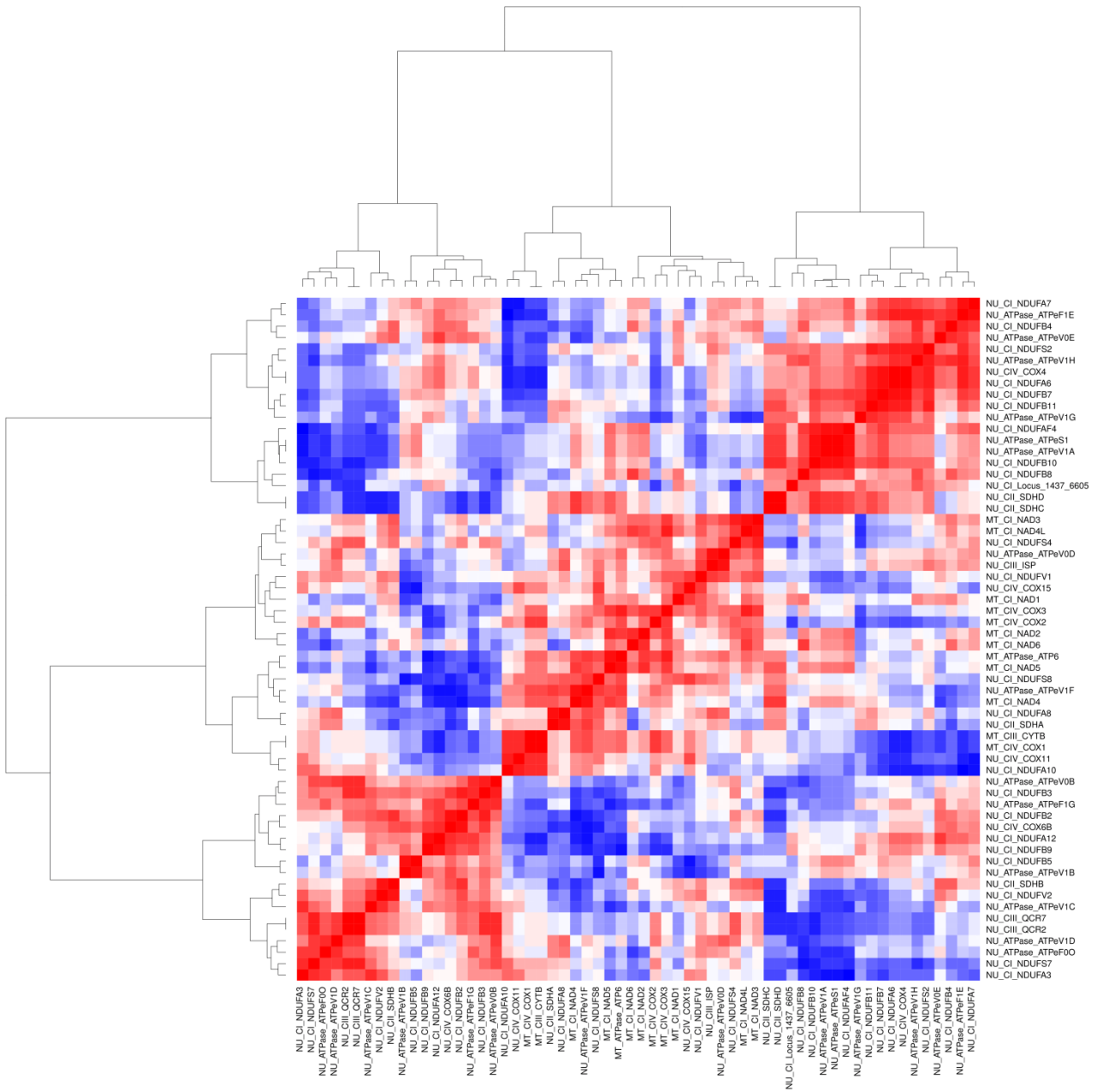
Supplementary figure 2: Transcription level of nuclear (NU) and mitochondrial (MT) OXPHOS subunits in *R. decussatus* (Rde) and *R. philippinarum* (Rph), separately for male (M) and female (F) samples.



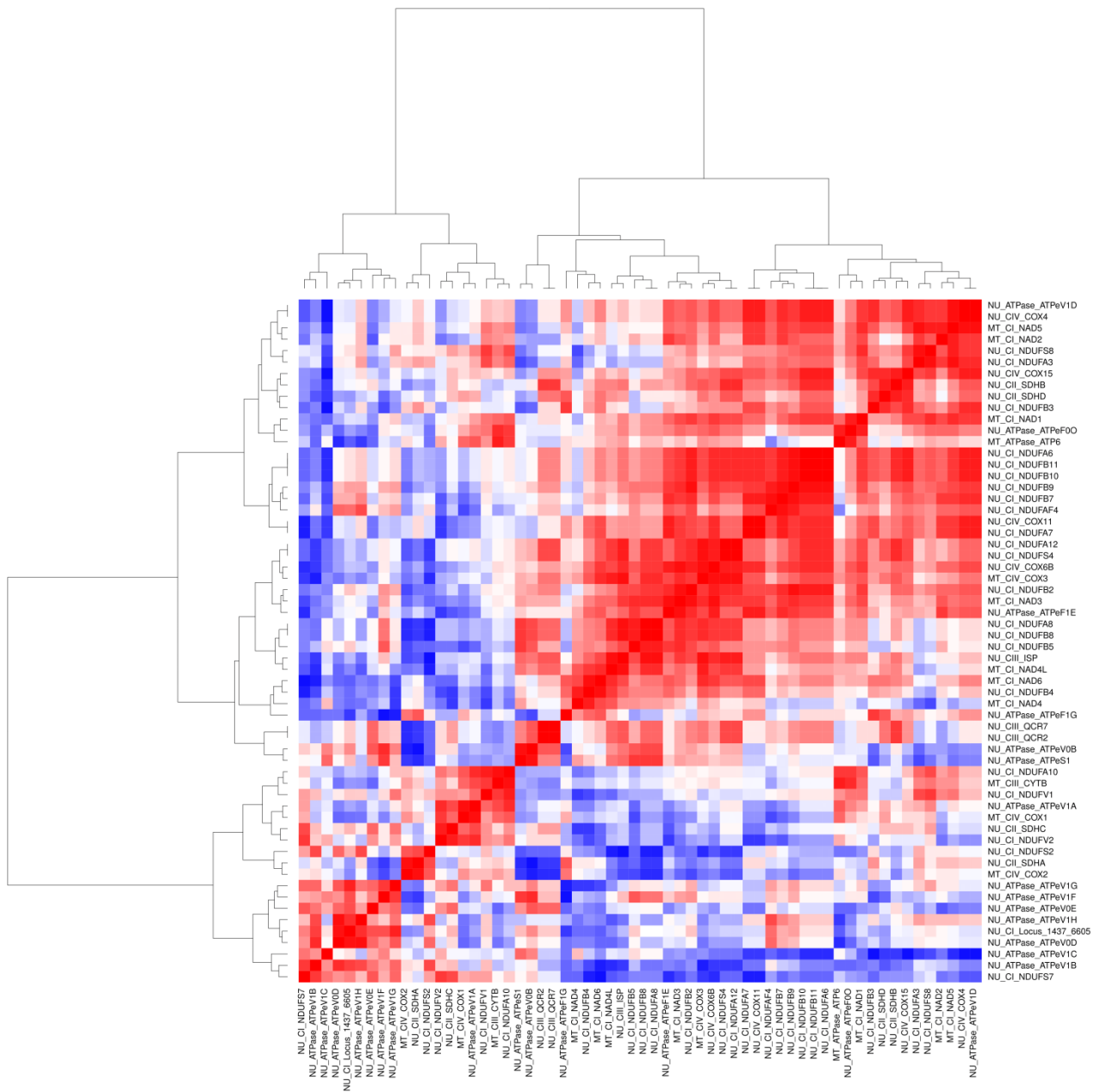
Supplementary figure 3: transcription level of nuclear (A) and mitochondrial (B) subunits separately for each OXPHOS complex in males of *R. decussatus*, females of *R. decussatus*, males of *R. philippinarum* and females of *R. philippinarum*.



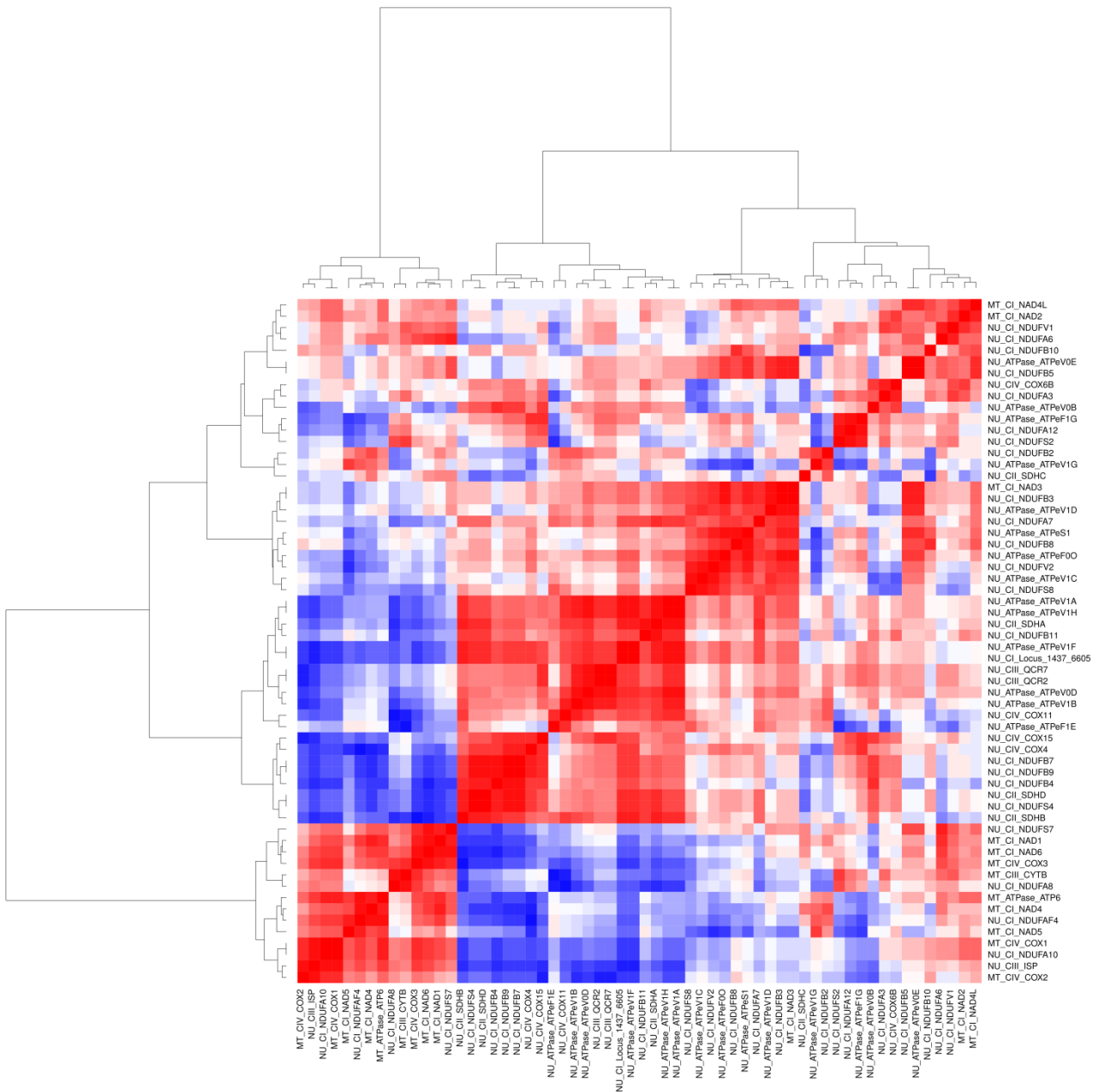
Supplementary figure 4: hierarchical clustering heatmap of OXPHOS subunits transcription in male (M1-M6) and female (F1-F6) samples of *R. decussatus* (Rde) and *R. philippinarum* (Rph).



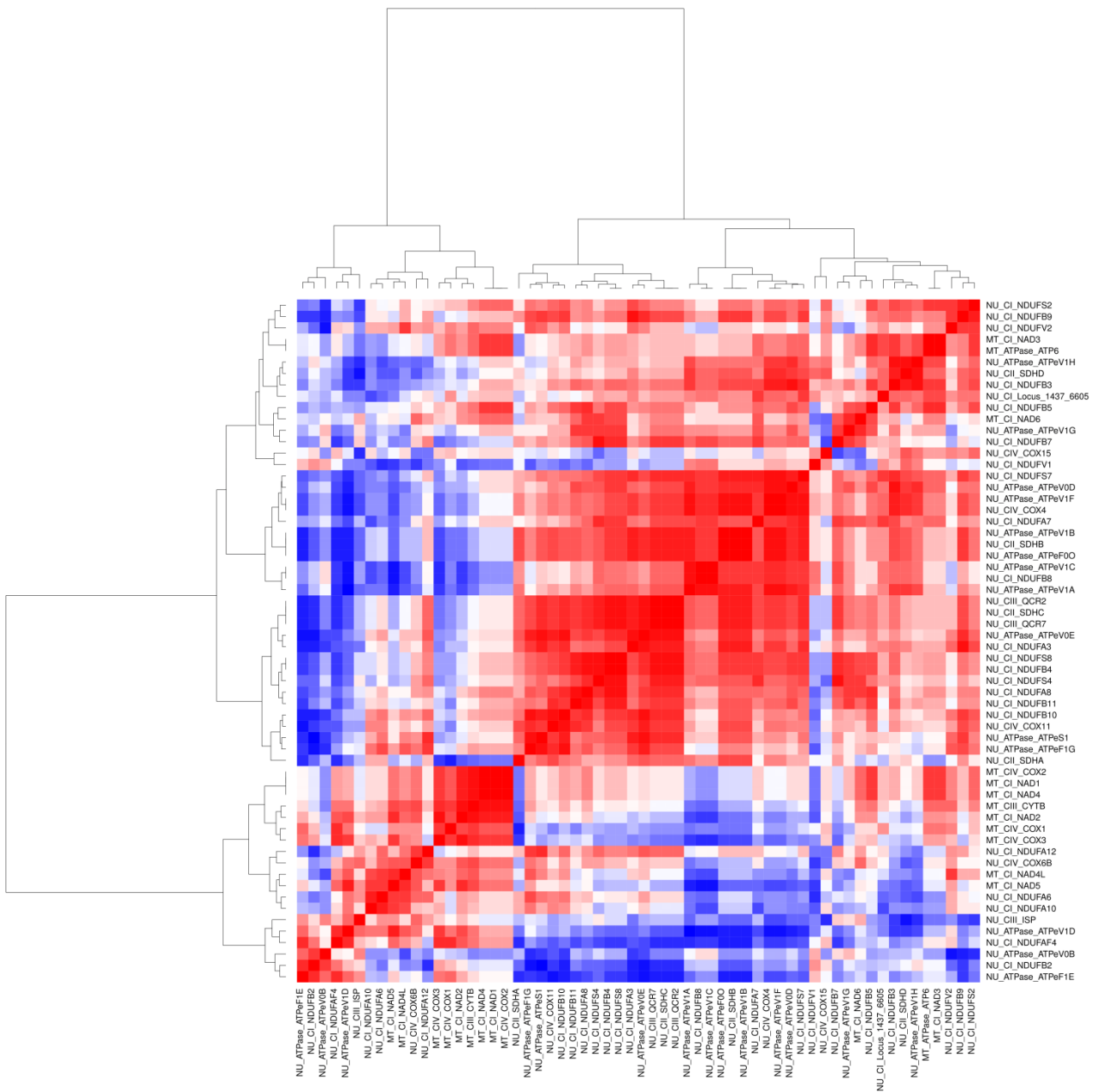
Supplementary figure 5: hierarchical clustering heatmap of transcription correlation matrices among OXPPOS subunits in males of *R. decussatus*. Red=positive correlation; Blue=negative correlation.



Supplementary figure 6: hierarchical clustering heatmap of transcription correlation matrices among OXPHOS subunits in females of *R. decussatus*. Red=positive correlation; Blue=negative correlation.



Supplementary figure 7: hierarchical clustering heatmap of transcription correlation matrices among OXPHOS subunits in males of *R. philippinarum*. Red=positive correlation; Blue=negative correlation.



Supplementary figure 8: hierarchical clustering heatmap of transcription correlation matrices among OXPHOS subunits in females of *R. philippinarum*. Red=positive correlation; Blue=negative correlation.

Supplementary table 1: Transcription levels (FPKM) of nuclear and mitochondrial OXPHOS subunits in males and females of *Ruditapes philippinarum*.

Name	Rph_F1	Rph_F2	Rph_F3	Rph_F4	Rph_F5	Rph_F6	Rph_M1	Rph_M2	Rph_M3	Rph_M4	Rph_M5	Rph_M6
NU_CI_NDUFS1	9.02843	2.42811	3.81139	10.5065	9.52189	10.2262	1.87598	6.14987	9.40763	14.4318	3.60722	11.2678
NU_CI_NDUFS2	2.48683	1.19202	4.05167	3.30141	6.00435	1.11809	1.79284	7.49446	3.32824	6.77676	4.05176	5.08449
NU_CI_NDUFS3	70.6618	95.6179	73.8358	114.322	103.659	101.656	107.343	140.473	104.764	147.512	119.658	147.581
NU_CI_NDUFS4	317.896	139.249	314.406	408.833	425.555	469.897	183.321	189.052	209.239	256.055	305.463	288.764
NU_CI_NDUFS7	173.334	139.77	198.798	163.827	131.669	131.523	116.05	130.256	125.688	199.349	143.12	173.876
NU_CI_NDUFS8	2.64739	6.45042	8.4926	10.1466	3.79971	8.73442	6.04826	19.484	11.224	22.5488	23.8286	22.7322
NU_CI_NDUFV1	9.64544	1.30209	5.8243	4.18471	4.51733	3.65911	6.96205	3.56333	4.33634	12.1489	0.812406	6.17088
NU_CI_NDUFV2	29.7933	38.2672	59.4522	66.735	43.0226	59.0929	15.8403	41.3371	31.6714	38.6636	28.8225	27.9752
NU_CI_NDUFA3	28.8745	19.8331	28.6337	22.8711	48.7834	25.9998	4.34527	22.4622	13.3739	23.8289	22.7783	22.3871
NU_CI_NDUFA4	57.5513	35.8771	31.6761	25.4076	15.0549	26.3694	17.1459	20.8691	39.138	14.0688	12.5103	14.2581
NU_CI_NDUFA7	75.7042	65.0396	108.544	122.439	65.9803	123.613	113.99	122.518	152.497	163.74	161.524	239.222
NU_CI_NDUFA8	3.35037	3.49741	7.40858	2.50695	5.4012	2.38793	2.12059	6.31793	4.48955	6.0769	11.7631	7.24271
NU_CI_NDUFA10	26.0626	7.33697	37.4311	0.007203	0.03657	0.65223	34.7889	55.1504	30.0882	0.977368	42.7296	0.677183
NU_CI_NDUFA12	5.90982	4.5129	14.9716	16.9245	19.8763	3.31754	1.46371	4.57194	5.43395	5.04106	7.02941	3.17635
NU_CI_NDUFB2	190.523	82.6986	77.8042	92.5717	68.0915	77.8454	123.735	75.0469	88.7854	71.9189	52.5672	99.3637
NU_CI_NDUFB3	178.706	110.351	217.328	228.843	172.426	211.925	314.409	367.607	312.235	369.01	349.109	433.754
NU_CI_NDUFB4	299.576	296.528	285.646	459.653	569.84	487.659	203.64	363.419	359.038	375.171	465.485	396.487
NU_CI_NDUFB5	212.717	130.669	250.775	239.477	191.256	203.746	285.47	423.032	393.596	412.474	412.628	494.445
NU_CI_NDUFB6	5.05955	0.07748	2.65158	1.49809	0.091832	0.047021	3.04437	8.63543	5.38312	2.8573	9.30637	0.023984
NU_CI_NDUFB7	37.972	19.5865	24.1134	40.5769	75.4173	60.1117	33.2672	33.1295	50.5741	52.0482	62.8846	62.5948
NU_CI_NDUFB9	19.5327	3.12404	17.6965	24.7321	30.1628	29.1613	19.0283	45.8051	22.6004	50.262	41.4162	30.9506
NU_CI_NDUFB10	10.2921	0.2015	19.4824	8.90959	12.8946	18.4385	10.7163	23.0875	16.2238	21.4978	26.675	20.0757
NU_CI_NDUFB11	126.61	67.7378	82.6335	124.786	123.359	135.562	162.66	191.468	154.476	179.286	259.705	228.358
NU_CII_SDHC	1.8131	3.01584	0.486838	4.1449	0.219107	0.10947	1.2757	3.12991	2.86094	6.4723	6.57852	3.80679
NU_CII_SDHD	52.9956	18.8094	29.6273	54.1351	60.1857	71.1792	51.6097	56.2047	20.6441	82.5868	50.9071	66.3728
NU_CII_SDHA	11.2953	0.87176	5.05843	11.5116	8.32319	12.3992	5.56921	4.96171	3.62208	7.58335	16.1058	3.97212
NU_CII_SDHB	1.12588	0.934187	0.65734	15.2964	5.76163	20.9904	0.791933	0.92778	0.70772	26.0714	4.10613	2.95079
NU_CIII_ISP	22.6858	34.5422	43.6189	0.551608	1.07632	3.05872	33.3434	6.83246	47.4173	0.687648	41.6776	7.86171
NU_CIII_QCR2	6.51899	1.56937	5.16311	7.76294	6.08646	5.2267	6.02446	7.74536	7.21305	16.2925	18.8457	15.6168

NU_CIII_QCR7	9.00852	27.9651	22.192	56.3005	44.9194	64.2702	39.9618	62.029	64.6775	100.959	111.223	158.479
NU_CIV_COX10	0	11.5619	9.21124	6.1152	8.20368	9.5855	4.73721	6.28552	19.5138	3.36148	2.8752	1.18388
NU_CIV_COX4	67.0821	13.8155	67.7196	84.5157	85.0554	79.6494	47.8632	50.4276	43.6096	97.1632	74.8571	80.8579
NU_CIV_COX6B	198.185	101.15	147.897	141.788	193.124	166.14	222.001	251.093	351.733	223.078	257.041	247.823
NU_CIV_COX11	34.21	19.3757	14.7921	37.7078	18.9221	34.3066	15.7142	31.5412	15.4274	30.6524	34.3189	25.5433
NU_CIV_COX15	15.8627	0.185439	2.68805	24.028	26.5659	6.34897	2.23894	3.89234	0.036828	2.2837	0.030677	1.9928
NU_ATPase_delta	278.236	108.562	261.982	329.433	207.955	342.446	179.247	298.616	245.997	398.389	356.087	264.705
NU_ATPase_epsilon	311.814	313.377	284.162	328.951	272.122	422.253	296.462	264.365	334.695	255.002	188.207	277.144
NU_ATPase_OSCP	295.826	100.572	356.22	370.613	330.832	355.172	150.529	177.432	125.796	379.653	304.061	278.041
NU_ATPase_b	17.911	15.7097	12.3675	14.8089	12.9775	7.83668	6.22236	11.0061	19.0371	9.91204	30.8882	19.8203
NU_ATPase_c	465.248	239.052	574.041	500.471	630.177	503.433	443.964	669.719	381.473	599.452	612.73	431.188
MT_CI_NAD1	3583.7	2512.83	3546.97	2406.53	2166.03	1748.42	3181.01	6207.81	4806.79	3768	4315.07	5104.98
MT_CI_NAD2	1146.48	416.69	658.49	418.85	434.54	516.15	790.13	2322.47	1881.88	778.66	1275.81	1390.21
MT_CI_NAD3	265.97	233.65	368.51	416.77	261.85	333.84	2182.84	6004.16	3492.29	4662.9	3105.79	7017.46
MT_CI_NAD4	6968.85	5753.12	4027.62	3547.56	3024.77	3375.32	1262.28	3418.75	2609.74	1470.32	1512.66	2613.76
MT_CI_NAD4L	259.91	134.86	325.81	224.45	152.98	230.09	1994.67	4311.35	3661.88	2399.95	2524.68	1784.26
MT_CI_NAD5	1333.28	1198.75	823.43	600.23	601.01	1084.12	1711.56	3088.26	2741.58	1030.5	2387.34	1411.81
MT_CI_NAD6	2670.08	1715.86	2873.24	1637.25	1504.9	1264.05	2662.28	4204	4497.43	3118.13	4515.03	8009.59
MT_CIII_CYTB	6258.28	5585.77	10761.08	5163.29	7628.8	5040.56	1536.19	6131.57	3744.7	2199.7	2357.58	3155.31
MT_CIV_COX1	16910.42	15760.12	20727.59	11849.75	12991.18	14795.3	4562.74	6670.17	7256.35	6067.6	4409.4	6494.71
MT_CIV_COX2	7408.05	7744.09	8686.58	4446.35	6151.67	7581.14	3453.15	8256.1	7298.01	5287.89	5863.8	7600.46
MT_CIV_COX3	11282.07	10867.87	12270.82	7543.64	7788.82	6037.4	4061.31	11201.07	8521.55	3163.32	3328.23	6522.49
MT_ATPase_ATP6	4144.65	2881.13	3206.93	1677.37	1536.73	1885.34	1785.02	5473.37	4450.6	5012.19	2938.4	6635.46

Supplementary table 2: Transcription levels (FPKM) of nuclear and mitochondrial OXPHOS subunits in males and females of *Ruditapes decussatus*.

Name	Rde_F1	Rde_F2	Rde_F3	Rde_F4	Rde_F5	Rde_F6	Rde_M1	Rde_M2	Rde_M3	Rde_M4	Rde_M5	Rde_M6
NU_CI_NDUFS1	0.256574	16.7106	3.33154	0	5.0892	9.18158	2.59195	0	6.46754	0	7.78973	0.198751
NU_CI_NDUFS2	42.5147	54.6407	5.26758	42.287	26.7237	44.0125	103.476	47.6604	58.6937	62.8251	67.4573	86.2929
NU_CI_NDUFS3	341.574	341.171	263.842	259.392	266.037	286.568	816.366	833.139	676.064	765.031	637.045	627.521
NU_CI_NDUFS4	386.802	165.185	214.523	246.135	240.983	240.627	648.893	680.014	834.103	666.53	524.091	445.37
NU_CI_NDUFS7	294.366	212.568	361.977	248.751	237.185	181.65	323.256	265.822	356.108	328.178	386.533	426.466
NU_CI_NDUFS8	167.27	246.767	160.411	187.073	192.321	148.41	521.313	460.639	540.737	428.78	402.236	517.306
NU_CI_NDUFV1	33.6196	28.2461	17.1711	21.5651	28.7509	7.31502	23.5278	20.6488	29.4908	15.7486	17.7176	32.2796
NU_CI_NDUFV2	40.1388	6.9801	13.2259	6.23965	13.5757	11.0722	7.52973	8.48713	10.1912	9.67821	9.32791	24.1668
NU_CI_NDUFA3	114.055	67.5886	116.964	60.6896	105.784	46.357	70.9505	42.3828	66.2619	36.3486	27.3013	48.0383
NU_CI_NDUFA4	89.4162	115.007	86.6172	122.289	104.028	139.8	100.073	73.6636	89.4358	56.3933	86.8765	53.5075
NU_CI_NDUFA7	153.77	132.651	96.4216	133.019	128.958	194.232	923.261	862.175	607.764	476.135	415.393	306.254
NU_CI_NDUFA8	80.9358	98.2589	80.6626	82.0531	66.5864	60.1945	88.6035	117.899	114.358	91.5422	93.8534	86.348
NU_CI_NDUFA10	313.717	358.263	469.629	397.088	425.566	189.254	463.147	481.533	582.719	397.057	385.279	673.996
NU_CI_NDUFA12	439.798	290.282	292.84	260.31	282.768	447.548	656.055	754.392	848.721	688.136	558.168	498.026
NU_CI_NDUFB2	485.688	305.396	432.744	330.634	362.583	466.313	1329.97	1378.07	1482.35	955.301	955.833	403.714
NU_CI_NDUFB3	535.608	245.096	414.251	361.062	329.662	413.064	651.101	507.287	515.774	523.301	379.792	443.614
NU_CI_NDUFB4	43.8565	31.5271	23.2771	29.3704	33.0757	123.523	224.033	270.96	194.067	227.714	222.031	56.8999
NU_CI_NDUFB5	2.94337	0.533173	4.6486	7.39293	1.0687	101.743	10.6282	670.381	65.2493	8.56799	290.369	1.83171
NU_CI_NDUFB6	77.4328	85.4307	56.4676	75.3248	43.7429	109.231	259.796	251.972	314.598	218.857	139.476	101.722
NU_CI_NDUFB7	46.5873	48.3753	42.7036	47.8067	41.406	75.5905	208.911	127.288	176.411	96.989	113.844	78.0546
NU_CI_NDUFB9	37.3713	25.524	46.1634	13.7877	17.0272	90.8604	206.817	198.337	329.196	158.161	184	58.8796
NU_CI_NDUFB10	144.642	280.89	160.132	244.643	176.39	380.787	1013.18	936.537	1040.23	780.003	723.49	558.245
NU_CI_NDUFB11	130.745	159.902	111.775	139.199	106.182	151.178	326.483	267.032	385.226	244.575	211.116	166.56
NU_CII_SDHC	0.806169	10.9878	5.01509	7.74532	6.13995	6.97471	6.52103	5.27481	16.8891	10.5562	7.23024	11.0986
NU_CII_SDHD	122.276	201.019	136.451	177.849	142.737	174.353	325.76	300.227	342.777	342.871	229.888	230.997
NU_CII_SDHA	80.5888	101.417	93.3534	96.3164	54.1203	31.5364	90.5438	49.9664	32.0673	54.6383	36.0965	71.596
NU_CII_SDHB	14.0585	0	0.017886	0	13.8131	0.229216	15.4906	0.705397	30.414	21.594	0.257981	0.120339
NU_CIII_ISP	43.8514	32.0455	1.47707	22.3062	21.8801	1.8049	38.1437	222.474	219.964	99.8521	59.9606	2.13514
NU_CIII_QCR2	44.6431	26.674	32.7279	32.6057	31.9574	31.5176	101.522	101.894	155.671	138.643	110.248	83.3897

NU_CIII_QCR7	570.15	364.729	367.685	333.784	355.353	667.961	1781.85	2056.1	2078.64	1723.57	1648.17	825.073
NU_CIV_COX10	512.095	159.731	364.629	167.416	251.732	88.4264	62.4519	65.5523	83.9875	65.2078	71.6373	53.9088
NU_CIV_COX4	554.097	654.531	427.868	536.512	422.091	733.219	2264.66	2003.77	2037.12	1922.5	1269.84	1644.16
NU_CIV_COX6B	850.122	480.145	704.339	509.467	576.891	1120.88	2498.85	2718.79	2671.8	2048.74	1844.39	1036.31
NU_CIV_COX11	1.86984	4.13135	8.0935	2.85903	8.97131	1.76175	3.34279	3.3148	2.30843	2.06289	1.42561	0.923233
NU_CIV_COX15	145.87	152.271	143.361	127.538	161.466	70.6341	44.1779	36.5379	61.4581	39.831	29.7333	34.6929
NU_ATPase_delta	511.765	874.829	527.858	725.927	610.509	656.873	1288.37	1090.88	1152.12	1142.9	964.526	1085.78
NU_ATPase_epsilon	572.94	541.086	396.65	427.979	407.959	1699.28	5916.87	6940.12	4742.68	4198.74	4482.52	1060.78
NU_ATPase_OSCP	515.682	378.333	484.933	335.729	326.427	270.316	594.452	957.07	681.824	563.802	547.596	602.187
NU_ATPase_b	66.6265	81.4617	57.1459	45.1624	43.2948	92.0064	163.492	131.068	166.154	184.502	93.6314	70.0394
NU_ATPase_c	2608.83	4005.07	2450.85	2995.39	2894.32	3762.9	7150.42	6721.86	6172.89	6179.16	5950.35	4442.8
MT_CI_NAD1	3135.269	3248.212	2330.334	2974.398	3863.442	3071.939	1565.915	1735.106	1899.751	1186.188	1117.561	1205.396
MT_CI_NAD2	2501.801	2280.796	2195.536	3162.572	2877.766	2542.762	1416.605	1284.135	1202.011	986.5388	1003.032	1070.944
MT_CI_NAD3	2212.241	1515.474	1096.656	2002.482	2032.695	1768.666	1637.737	1784.499	1751.381	865.6536	1241.233	614.7322
MT_CI_NAD4	2899.706	5175.253	4056.368	7216.748	5159.5	4056.14	2755.904	2868.544	2160.515	2812.234	2773.395	2628.979
MT_CI_NAD4L	1692.393	1009.684	974.7059	1321.445	1750.217	1182.136	993.1324	1815.467	1262.647	1499.61	1078.721	983.25
MT_CI_NAD5	1548.966	2150.238	1375.886	2624.303	2205.046	1994.498	1320.87	1083.028	1172.797	836.3867	766.1857	1072.841
MT_CI_NAD6	3319.512	2933.951	2056.86	3360.866	4160.5	3791.407	4334.672	5196.441	4092.202	4734.026	3400.761	1623.972
MT_CIII_CYTB	5712.202	6297.558	6385.257	7334.151	8406.908	4111.683	1941.87	2134.804	2022.261	1879.716	1575.928	2505.89
MT_CIV_COX1	6139.872	6385.616	7301.02	8237.131	9016.431	5398.148	3378.831	3449.165	3480.555	3843.588	3058.105	4378.005
MT_CIV_COX2	5791.332	5004.019	5612.452	6511.007	6489.742	3844.308	4005.916	3673.058	3377.702	3753.142	3619.641	4249.544
MT_CIV_COX3	5572.661	4506.484	3920.555	5830.732	6517.493	4088.493	4009.083	4316.274	4229.351	4041.913	3475.111	3471.22
MT_ATPase_ATP6	3577.745	3603.098	2222.948	4353.33	3952.429	2833.674	1656.353	2196.444	1818.909	1769.062	1395.27	2065.612

Supplementary table 3: Top) Medians of correlations between : i) nuclear subunits; ii) mitochondrial subunits; iii) nuclear versus mitochondrial subunits; iiiii) all subunits; Bottom) Medians of correlations between nuclear subunits of CI, CII, CIII, CIV, CV

	Nuclear subunits	Mitochondrial subunits	Nuclear ~ mitochondrial subunits	All
<i>R. decussatus</i> M	0.14	0.28	0.08	0.25
<i>R. decussatus</i> F	0.07	0.6	-0.08	0.14
<i>R. philippinarum</i> M	0.37	0.65	0.08	0.14
<i>R. philippinarum</i> F	0.25	0.54	-0.08	0.03

	CI	CII	CIII	CIV	CV
<i>R. decussatus</i> M	0.37	0.25	0.82	0.85	0.14
<i>R. decussatus</i> F	0.08	0.48	0.54	-0.11	0.14
<i>R. philippinarum</i> M	0.37	0.68	0.37	0.14	0.42
<i>R. philippinarum</i> F	0.2	0.85	0.11	0.4	0.42




Supplementary table 4: medians of dN/dS for each OXPHOS complex. NU Rde~Rph= dN/dS between nuclear subunits of *Ruditapes decussatus* and *Ruditapes philippinarum*; MT Rde~Rph F=dN/dS between mitochondrial subunits of *Ruditapes decussatus* and *Ruditapes philippinarum* F type; MT Rde~Rph M=dN/dS between mitochondrial subunits of *Ruditapes decussatus* and *Ruditapes philippinarum* M type.

	CI	CII	CIII	CIV	ATPase
NU Rde~Rph	0.16	0.04	0.28	0.09	0.14
	CI	CIII	CIV	ATPase	
MT Rde~Rph F	0.08	0.08	0.15	0.19	
MT Rde~Rph M	0.12	0.08	0.18	0.18	

Supplementary table 5: most abundant GO annotations associated to genes with mitochondrial annotations and dN/dS > 0.2 .

GO term	Description	%
GO:0006810	transport	17,62%
GO:0044281	small molecule metabolic process	15,14%
GO:0055114	oxidation-reduction process	15,06%
GO:0006351	transcription, DNA-templated	10,66%
GO:0006355	regulation of transcription, DNA-templated	9,92%
GO:0055085	transmembrane transport	8,92%
GO:0016310	phosphorylation	7,76%
GO:0006412	translation	5,69%
GO:0006811	ion transport	5,34%
GO:0006468	protein phosphorylation	4,14%
GO:0035556	intracellular signal transduction	4,00%
GO:0032259	methylation	3,10%
GO:0008652	cellular amino acid biosynthetic process	2,93%
GO:0006974	cellular response to DNA damage stimulus	2,36%
GO:0015031	protein transport	2,25%
GO:0006281	DNA repair	2,23%
GO:0006310	DNA recombination	1,64%
GO:0006260	DNA replication	1,58%

Supplementary table 6: mitochondrial GO annotation of *R. decussatus* and *R. philippinarum* specific genes

 shared name	 rde_present	 rph_present
transcription from mitochondrial promoter	<input checked="" type="checkbox"/>	<input type="checkbox"/>
sperm mitochondrion organization	<input type="checkbox"/>	<input checked="" type="checkbox"/>
regulation of mitochondrion organization	<input checked="" type="checkbox"/>	<input type="checkbox"/>
regulation of mitochondrial membrane permeability	<input checked="" type="checkbox"/>	<input type="checkbox"/>
rRNA import into mitochondrion	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
protein localization to mitochondrion	<input type="checkbox"/>	<input checked="" type="checkbox"/>
positive regulation of mitochondrial fission	<input checked="" type="checkbox"/>	<input type="checkbox"/>
positive regulation of mitochondrial calcium ion concentration	<input checked="" type="checkbox"/>	<input type="checkbox"/>
negative regulation of release of cytochrome c from mitochondria	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrion transport along microtubule	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrion organization	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrion morphogenesis	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrion localization	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial unfolded protein response	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial transport	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial translational termination	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial translational elongation	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial respiratory chain complex IV assembly	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial respiratory chain complex I assembly	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial protein catabolism	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial ncRNA surveillance	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial genome maintenance	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial fusion	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial fragmentation involved in apoptotic process	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial fission	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial electron transport, ubiquinol to cytochrome c	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial calcium uptake	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial calcium ion transport	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial calcium ion homeostasis	<input checked="" type="checkbox"/>	<input type="checkbox"/>
mitochondrial RNA 3'-end processing	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mitochondrial DNA replication	<input checked="" type="checkbox"/>	<input type="checkbox"/>
intracellular distribution of mitochondria	<input checked="" type="checkbox"/>	<input type="checkbox"/>
inner mitochondrial membrane organization	<input type="checkbox"/>	<input checked="" type="checkbox"/>
apoptotic mitochondrial changes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Conclusions

The work carried out in this thesis allowed me to develop some methodological approaches for HTS data analysis to be used in a comparative framework. I was able to investigate the evolution of nuclear and mitochondrial genes in two mollusc species—*Ruditapes decussatus* and *Ruditapes philippinarum* (Bivalvia, Veneridae)—providing insights into biological processes in a Phylum where genomics data are largely missing. Compared to taxa investigated so far, the study of these two non-model species offered different perspectives on several biological issues.

By analyzing RNA-seq from gonads of *R. decussatus* and *R. philippinarum*, I investigated—for the first time in Mollusca—the transcription level and sequence evolution of sex-biased genes. These species showed a low number of sex-biased genes, compared to other taxa. This is not surprising since clams lack sexual dimorphism as well as mating behavior, which are responsible for the majority of differential transcription between sexes. Nevertheless, there are considerable differences in the transcription of such genes: in 86% of orthologs, the sex-bias is not maintained between the two species and the most frequent condition is represented by genes that are female-biased in one species and unbiased in the other species. This is surprising, since male-biased genes show the greater transcription divergence in most studies. On the other hand, I investigated the sequence evolution of sex-biased genes and I found that genes that are male-biased in both the species are characterized by a higher rate of protein evolution, confirming a pattern observed in most of the taxa. This evidence suggests that both transcription level divergence and rate of protein sequence variation act on the evolution of sex-biased genes. In particular, a faster evolution of protein sequence seems to be predominant for male-biased genes in the analyzed species, while a more variable transcriptional regulation is observed for female-biased genes, suggesting that these two kinds of evolution may be decoupled. In addition, the investigation of all genes in both the species revealed a lack of correlation between rate of protein evolution and

transcription level. Since the evolutionary rate of a protein would be mainly influenced by its transcription level according to recentest theories, the lack of correlation in these two bivalve species suggest that this hypothesis might not account for coding sequence evolution in all the Metazoa.

The study of bivalve species offered also an exceptional opportunity to investigate the dynamics of mito-nuclear evolution. In fact, *R. decussatus* and *R. philippinarum* are characterized by different mechanisms of mitochondrial inheritance, and *R. philippinarum* presents a natural condition of mitochondrial heteroplasmy. In the work reported in Chapter 4, I took advantage of this condition to infer the co-evolution of mitochondrial and nuclear genome. I found that in these species the rate of protein evolution (measured as dN/dS) of mitochondrial OXPHOS subunits is an order of magnitude higher compared to that in species where mito-nuclear coevolution was investigated so far; on the contrary dN/dS of nuclear OXPHOS subunits is comparable to that in other species. This finding does not support one of the main hypothesis about the dynamics of mito-nuclear coevolution, the 'nuclear compensation hypothesis', which posits that nuclear subunits evolve faster to compensate the high mutation level acting on mtDNA, thus ensuring the proper molecular interaction among OXPHOS subunits. Although nuclear-encoded OXPHOS subunits do not seem to be affected by mitochondrial variability, I found that other nuclear genes could be likely involved in mito-nuclear incompatibilities. Genes involved in regulation of transcription, translation, and recombination of mitochondrial DNA, as well as genes for the respiratory chain assembly and response to oxidative stress were characterized by significantly higher dN/dS in the analyzed species, suggesting a fast evolution in response to the high mitochondrial protein sequence variability.

If, on one hand, the purpose of this thesis was to increase the knowledge about molluscs, on the other hand most biological issues remain unexplored in these species and little is known about their evolution, adaptation and development. Genomics data are now, more than ever, easy to obtain. Nevertheless, the analysis of such complex and extensive data is still not optimized for non-model organisms. During this work, I indeed had to face several technical issues: most of the bioinformatic tools for inferring differential expression analysis or enrichment of

biological processes were discarded, since they worked only on model species. In addition, I had to develop an annotation pipeline for non-model organisms that combined different tools and databases in order to detect remote homology and be non-model organism-friendly. Finally, the detection of orthology and paralogy have remarkable limitations when distantly-related species are investigated, and this can have deep impact in the interpretation of results. Comparative genomics has a fundamental role in almost every field of biology, so an improvement in comparative methods—both theoretical and technical—is a key point for a better exploitation of genomics data in the next future.

References

Albertin CB et al. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*. 524:220–224.

Altschul SF et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Amemiya CT et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature*. 496:311–316.

Ashburner M et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25:25–29.

Birky CW Jr. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc. Natl. Acad. Sci. U. S. A.* 92:11331–11338.

Boothby TC et al. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U. S. A.* 112:15976–15981.

Breton S et al. 2014. A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. *Trends Genet.* 30:555–564.

Breton S et al. 2009. Comparative mitochondrial genomics of freshwater mussels (Bivalvia: Unionoida) with doubly uniparental inheritance of mtDNA: gender-specific open reading frames and putative origins of replication. *Genetics*. 183:1575–1589.

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34:D363–8.

Chiari Y, Cahais V, Galtier N, Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* 10:65.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437:69–87.

Doucet-Beaupré H et al. 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.

Dunn CW, Ryan JF. 2015. The evolution of animal genomes. *Curr. Opin. Genet. Dev.*

35:25–32.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29:51–63.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.* 8:689–698.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

Falda M et al. 2012. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics.* 13 Suppl 4:S14.

Finn RD et al. 2014. Pfam: the protein families database. *Nucleic Acids Res.* 42:D222–30.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39:W29–37.

Fischer S et al. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinformatics.* Chapter 6:Unit 6.12.1–19.

da Fonseca RR et al. 2016. Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Mar. Genomics.* 30:3–13.

Ghiselli F et al. 2012. De Novo assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. *Mol. Biol. Evol.* 29:771–786.

Ghiselli F et al. 2013. Structure, transcription, and variability of metazoan mitochondrial genome: perspectives from an unusual mitochondrial inheritance system. *Genome Biol. Evol.* 5:1535–1554.

Ghiselli F et al. 2017. The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae). *PeerJ.* 5:e3692.

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.

GIGA Community of Scientists et al. 2014. The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *J. Hered.* 105:1–18.

Gosling EM. 2003. *Bivalve Molluscs: Biology, Ecology and Culture*. Hoboken: Wiley Online Library.

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.
- Harrison PW, Wright AE, Mank JE. 2012. The evolution of gene expression and the transcriptome-phenotype relationship. *Semin. Cell Dev. Biol.* 23:222–229.
- Haszprunar G, Schander C, Halanych KM. 2008. In *Phylogeny and Evolution of the Mollusca* (eds Ponder, W. & Lindberg, D. R.) 19–32 (Univ. of California Press)
- Haszprunar G, Wanninger A. 2012. *Molluscs. Curr. Biol.* 22:R510–4.
- Hochner B, Glanzman DL. 2016. Evolution of highly diverse forms of behavior in molluscs. *Curr. Biol.* 26:R965–R971.
- Jones P et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 30:1236–1240.
- Koepfli K-P, Paten B, Genome 10K Community of Scientists, O'Brien SJ. 2015. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci.* 3:57–111.
- Koutsovoulos G et al. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc. Natl. Acad. Sci. U. S. A.* 113:5053–5058.
- Krasileva KV et al. 2013. Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* 14:R66.
- Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* 4:237.
- Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Res.* 6. doi: 10.12688/f1000research.12232.1.
- Lane N. 2012. The problem with mixing mitochondria. *Cell.* 151:246–248.
- Lechner M et al. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics.* 12:124.
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–2189.
- Merchant S, Wood DE, Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ.* 2:e675.
- Milani L, Ghiselli F, Guerra D, Breton S, Passamonti M. 2013. A comparative analysis of mitochondrial ORFans: new clues on their origin and role in species with doubly uniparental inheritance of mitochondria. *Genome Biol. Evol.* 5:1408–1434.
- Milani L, Ghiselli F, Iannello M, Passamonti M. 2014. Evidence for somatic transcription of male-transmitted mitochondrial genome in the DUI species *Ruditapes*

philippinarum (Bivalvia: Veneridae). *Curr. Genet.* 60:163–173.

Milani L, Ghiselli F, Maurizii MG, Nuzhdin SV, Passamonti M. 2014. Paternally transmitted mitochondria express a new gene of potential viral origin. *Genome Biol. Evol.* 6:391–405.

Milani L, Ghiselli F, Passamonti M. 2016. Mitochondrial selfish elements and the evolution of biological novelties. *Curr. Zool.* 62:687–697.

Moroz LL et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. *Nature.* 510:109–114.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 5:621–628.

Nystedt B et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature.* 497:579.

Ogura A, Ikeo K, Gojobori T. 2004. Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human. *Genome Res.* 14:1555–1561.

Parker J, Helmstetter AJ, Devey D, Wilkinson T, Papadopoulos AST. 2017. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci. Rep.* 7:8345.

Passamonti M, Ghiselli F. 2009. Doubly uniparental inheritance: two mitochondrial genomes, one precious model for organelle DNA inheritance and evolution. *DNA Cell Biol.* 28:79–89.

Passamonti M, Ricci A, Milani L, Ghiselli F. 2011. Mitochondrial genomes and Doubly Uniparental Inheritance: new insights from *Musculista senhousia* sex-linked mitochondrial DNAs (Bivalvia Mytilidae). *BMC Genomics.* 12:442.

Passamonti M, Scali V. 2001. Gender-associated mitochondrial DNA heteroplasmy in the venerid clam *Tapes philippinarum* (Mollusca Bivalvia). *Curr. Genet.* 39:117–124.

Qiu Q et al. 2012. The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44:946.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16:276–277.

Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* 13:505–516.

Sato M, Sato K. 2013. Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. *Biochim. Biophys. Acta.* 1833:1979–1984.

Schell T et al. 2017. An annotated draft genome for *Radix auricularia* (Gastropoda,

Mollusca). bioRxiv. 087254. doi: 10.1101/087254.

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.

Smith SA et al. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*. 480:364–367.

Tamvacakis AN, Senatore A, Katz PS. 2015. Identification of genes related to learning and memory in the brain transcriptome of the mollusc, *Hermisenda crassicornis*. *Learn. Mem.* 22:617–621.

The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*. 487:94.

Voolstra CR, Giga Community of, Wörheide G, Lopez JV. 2017. Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA). *Invertebr. Syst.* 31:1–7.

Zhan X et al. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat. Genet.* 45:563.

Zhang G et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*. 490:49–54.

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