A Comparative Analysis of Mitochondrial ORFans: New Clues on Their Origin and Role in Species with Doubly Uniparental Inheritance of Mitochondria

Liliana Milani[1,*], Fabrizio Ghiselli[1], Davide Guerra[1], Sophie Breton[2], and Marco Passamonti[1]

[1]Dipartimento di Scienze Biologiche, Geologiche ed Ambientali, University of Bologna, Bologna, Italy

[2]Département de Sciences Biologiques, Université de Montréal, Montréal, Québec, Canada

*Corresponding author: E-mail: liliana.milani@unibo.it.

Accepted: June 27, 2013

Abstract

Despite numerous comparative mitochondrial genomics studies revealing that animal mitochondrial genomes are highly conserved in terms of gene content, supplementary genes are sometimes found, often arising from gene duplication. Mitochondrial ORFans (ORFs having no detectable homology and unknown function) were found in bivalve molluscs with Doubly Uniparental Inheritance (DUI) of mitochondria. In DUI animals, two mitochondrial lineages are present: one transmitted through females (F-type) and the other through males (M-type), each showing a specific and conserved ORF. The analysis of 34 mitochondrial major Unassigned Regions of *Musculista senhousia* F- and M-mtDNA allowed us to verify the presence of novel mitochondrial ORFs in this species and to compare them with ORFs from other species with ascertained DUI, with other bivalves and with animals showing new mitochondrial elements. Overall, 17 ORFans from nine species were analyzed for structure and function. Many clues suggest that the analyzed ORFans arose from endogenization of viral genes. The co-option of such novel genes by viral hosts may have determined some evolutionary aspects of host life cycle, possibly involving mitochondria. The structure similarity of DUI ORFans within evolutionary lineages may also indicate that they originated from independent events. If these novel ORFs are in some way linked to DUI establishment, a multiple origin of DUI has to be considered. These putative proteins may have a role in the maintenance of sperm mitochondria during embryo development, possibly masking them from the degradation processes that normally affect sperm mitochondria in species with strictly maternal inheritance.

Key words: mitochondrial ORFans, mitochondrial inheritance, Doubly Uniparental Inheritance of mitochondria, endogenous virus.

Introduction

Comparative mitochondrial genomics revealed that animal mitochondrial DNAs (mtDNAs) are highly conserved in terms of gene content (Boore 1999; Gissi et al. 2008). These small, typically circular and intron-less molecules encode 2 ribosomal RNAs, 22 transfer RNAs, and 13 protein subunits of the mitochondrial respiratory complexes and ATP synthase. The other subunits of the electron transport chain and all the proteins involved in other mitochondrial functions, such as mtDNA replication and expression, are encoded by the nucleus (Boore 1999). However, supplementary genes are sometimes found in mtDNA. Many mechanisms are responsible for the origin of such new genes. For example, novel mitochondrial Open Reading Frames (ORFs) can arise from gene duplication. In bivalve molluscs, a *cox2* duplication is found in the clam *Ruditapes philippinarum* (Bivalvia, Veneridae) (Okazaki M and Ueshima R, unpublished data; GenBank AB065375.1) and in the mussel *Musculista senhousia* (Bivalvia, Mytilidae) (Passamonti et al. 2011). Moreover, *nad2* duplication is at the origin of two novel ORFs in the oyster genus *Crassostrea* (Bivalvia, Ostreidae) (Wu et al. 2012). Extra elements were also found in Cnidaria mtDNA, either from duplication of extant genes or not: a duplicated *cox1* in some hydroidolinan hydrozoans (Cnidaria, Hydrozoa), two novel ORFs in Medusozoa (Kayal et al. 2011), and a novel ORF in every octocoral (Cnidaria, Anthozoa) that has been screened to date (McFadden et al. 2010). One of the two medusozoan ORFs shares several conserved motifs characteristic of the

polymerase domain typical of family B-DNA polymerases (polB; Shao et al. 2006). The other ORF, named ORF314, do not resemble any other known protein. Kayal et al. (2011) attributed the origin of these two extra elements to an ancient invasion by a linear plasmid that caused the linearization of the mtDNA in Medusozoa, consistent with a previously established hypothesis for polB-like sequences found in the linear mtDNA of fungi and algae (Mouhamadou et al. 2004). The conservation of both sequence length and position suggested some level of selection pressure for their maintenance in the mtDNA of most medusozoans (Kayal et al. 2011). The octocoral extra ORF is recognized as a putatively DNA mismatch repair protein (mtMutS) (Pont-Kingdon et al. 1995; Claverie et al. 2009; Bilewitch and Degnan 2011; Ogata et al. 2011). As for medusozoan ORFs, mtMutS was supposed to be originated by horizontal gene transfer, but in this case either through an epsilonproteobacterium or a viral infection (Claverie et al. 2009; Bilewitch and Degnan 2011; Ogata et al. 2011).

Interestingly, novel mitochondrial ORFs have been also discovered in bivalve molluscs with Doubly Uniparental Inheritance (DUI) of mitochondria (Skibinski et al. 1994a, 1994b; Zouros et al. 1994a, 1994b). Specifically, in metazoans, mitochondria are commonly inherited maternally by Strictly Maternal Inheritance (SMI) (Birky 2001), whereas in DUI animals two mitochondrial lineages are present: one transmitted through females (F-type) and the other through males (M-type). In DUI bivalves, females inherit F-type mtDNA, whereas males inherit both F- and M-types (Skibinski et al. 1994a, 1994b; Zouros et al. 1994a, 1994b). In DUI bivalves (orders Mytiloida, Unionoida, and Veneroida), two novel lineage-specific ORFs were found, one in the F-mtDNA (fORF) and one in the M-mtDNA (mORF) (Breton et al. 2009; Breton et al. 2011a, 2011b; Ghiselli et al. 2013). These novel ORFs have been hypothesized to be responsible for the different mode of mtDNA transmission and the maintenance of gonochorism in DUI bivalves (Breton et al. 2009, 2011a, 2011b).

In all the analyzed DUI *Mytilus* species, the novel fORF is localized in the Largest Unassigned Region (LUR) and encodes a putative protein of more than 100 amino acids (aa), suggesting its maintenance in the subfamily Mytilinae for more than 10 million years (Breton et al. 2011b). A fORF is present also in the F-mtDNA of *Musculista senhousia*, a DUI mytilid of the subfamily Crenellinae (Breton et al. 2011b). In the venerid *R. philippinarum*, the fORF is localized in the Female Largest Unassigned Region (FLUR), whereas the mORF in the Male Unassigned Region 21 (Ghiselli et al. 2013). Interestingly, the two lineage-specific ORFs found in the freshwater mussel *Venustaconcha ellipsiformis* (Bivalvia, Unionidae), the fORF (found between *tRNA-Glu* and *nad2*) and the mORF (found between *tRNA-Asp* and *nad4L*), are both translated (Breton et al. 2009), and the female-transmitted novel protein is not only present in mitochondria but also in the nuclear membrane and in egg nucleoplasm (Breton et al. 2011a). These findings might support an involvement of these novel mitochondrial genes in some, still unknown, key biological functions in bivalve species with DUI. For instance, it has been suggested that the newly identified mtORFs in DUI bivalves might have a role in determining the fate of sperm mitochondria in fertilized eggs, maybe leading to the two distribution patterns of spermatozoon mitochondria observed in DUI early embryos: the aggregated pattern, in which these mitochondria form a cluster along the cleavage furrow in two-blastomere embryos and among blastomeres in four-cell embryos, and the dispersed pattern, in which sperm mitochondria are randomly scattered (Cao et al. 2004; Cogswell et al. 2006; Milani et al. 2011, 2012).

The analysis of 34 mitochondrial major Unassigned Regions (URs) of *M. senhousia* F- and M-mtDNA allowed us to verify the presence of novel mitochondrial ORFs in this species and to compare them with novel ORFs from other bivalve species with ascertained DUI, with other bivalves and with animals showing new mitochondrial elements. We found that many features are shared by all novel ORFs, allowing us to formulate an hypothesis on their possible shared origin.

Materials and Methods

Gametes Collection, DNA Extraction, PCRs, and Sequencing

M. senhousia specimens from Venice lagoon (Italy) were induced to spawn in sea water with oxygen peroxide, according to Morse et al. (1977). Each spawning was analyzed with a light microscope to sex specimens. Sperm and eggs were collected and then centrifuged at 3,000 × g; after that, sea water was removed and replaced with ethanol. Gametes were stored at −20 °C. Total DNA extraction from gametes of 11 females and 12 males was performed with DNeasy Tissue Kit (Qiagen) following manufacturer instructions. All polymerase chain reactions (PCRs) were executed on a 2720 Thermal Cycler (Applied Biosystems). All primers were provided by Invitrogen™ (see list of primers in supplementary material S1, Supplementary Material online).

Long PCRs, using gamete DNA extractions as template, were performed to obtain a segment containing the whole Largest Unassigned Region (LUR) (i.e., in both mtDNAs, the region between *rrnL* and *cob*); in the F-mtDNA, this region also contains the Female Unassigned Region 2 (FUR2) (see Passamonti et al. 2011 for annotation details). Primers for long-PCRs are the same used in Passamonti et al. (2011): M-mtDNA from sperm was amplified with primers M-16S103F and M-cob386R, whereas F-mtDNA from eggs with primers F-16S142F and F-cob383R (supplementary material S1, Supplementary Material online). Both segments were amplified with Herculase II Fusion Enzyme kit (Stratagene) in a 50 µl reaction volume composed of 10 µl

5× Herculase II Run Buffer, 0.5 μl of 100 mM dNTP mix, 1.25 μl of 10 μM primers, 0.5 μl of Herculase II Fusion DNA Polymerase, 5 μl of total DNA, and 31.5 μl of Nuclease-free water (Ambion Inc.). Long PCR cycles followed the same scheme for the M- and the F-mtDNA. The reactions started with an initial denaturation at 95 °C for 5 min, then 30 cycles of denaturation at 95 °C for 20 s, annealing at 48 °C for 20 s and extension at 68 °C for 10 s, then a final extension at 68 °C for 8 min.

Long PCR products were used as a template to amplify single overlapping segments of the LURs and the FUR2 with standard PCRs. Primers for standard PCRs (supplementary material S1, Supplementary Material online) were designed with Primer3 (Rozen and Skaletsky 2000) on the two complete *M. senhousia* F- and M-mtDNAs (GenBank accession nos. GU001953–4). GoTaq® Flexi Dna Polymerase (Promega) kit was used for standard PCRs. Reactions were performed in a 50 μl volume composed of 10 μl of 5× Green GoTaq Flexi Buffer, 6 μl of 25 mM $MgCl_2$, 1 μl of 40 μM dNTP mix (10 μM each dNTP), 2.5 μl of 10 μm primers, 0.25 μl of GoTaq Dna Polymerase 5 U/μl, 4 μl of template DNA from the long PCRs, and 24 μl of Nuclease-free water (Ambion Inc.). LURs and FUR2 were amplified with the following cycle: initial denaturation at 95 °C for 2 min, 30 cycles of denaturation at 95 °C for 30 s, annealing at 48 °C for 30 s, extension at 72 °C for 90 s, and a final extension at 72 °C for 5 min.

All PCR products were purified with Wizard SV Gel and PCR clean-up System (Promega) kit, GenElute PCR clean-up kit, and GenElute Extraction kit (Sigma-Aldrich), following manufacturer instructions. Sequencing was performed at Macrogen Inc. (Seoul, South Korea). Sequences were assembled and aligned with MEGA5 (Tamura et al. 2011).

Novel Mitochondrial ORFs

Nucleotide Level: Sequence Conservation

We used ORF Finder (http://www.ncbi.nlm.nih.gov/gorf, last accessed July 23, 2013) to assess the presence of novel ORFs in DUI species LURs present in GenBank, using the invertebrate mitochondrial genetic code. For DUI species, novel mitochondrial sex-specific ORFs were already described and confirmed in literature (*Mytilus* spp., *M. senhousia*: Breton et al. 2011b; *V. ellipsiformis*: Breton et al. 2009; *R. philippinarum*: Ghiselli et al. 2013). The obtained sequences of *M. senhousia* FUR2 and 689 annotated mt LURs of four *Mytilus* species (*Mytilus californianus*, *Myt. edulis*, *Myt. galloprovincialis*, and *Myt. trossulus*) (Bivalvia, Mytilidae) were checked to assess the conservation of the ORFs described in Passamonti et al. (2011) and Breton et al. (2011b) (last GenBank access: September 2012). The new sequences of *M. senhousia* LURs were also searched for the presence of novel ORFs (only the longest ORFs found in all sequences were considered). In the analyzed DUI species, we will refer to the ORFs present either in the F or the M mtDNA (i.e., lineage-specific ORFs) as fORF and mORF, respectively. For comparison, ORFs were searched also in the LUR of the venerid *Paphia euglypta*, a species in which the presence of DUI has not been investigated yet (only one LUR sequence is available; table 1). Specific names are given to non-lineage-specific extra mtORFs, comprising mtORFs in non-DUI species. p-distances of novel ORFs of *M. senhousia* and other DUI species were calculated with MEGA5 (Tamura et al. 2011) using the bootstrap method on all suitable sequences available in GenBank.

Protein Level: Structural and Functional Analysis

The above-mentioned ORFs were translated and analyzed at the amino acid level (see table 1 for the sequences in which the analyzed ORFs are included, and supplementary material S2, Supplementary Material online, for amino acid sequences). We will refer to the translations of fORFs and mORFs of DUI species as FORF and MORF, respectively.

To find Signal Peptides (SPs) we used Phobius (http://phobius.sbc.su.se/, last accessed July 23, 2013; Käll et al. 2004), InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/, last accessed July 23, 2013; Zdobnov and Apweiler 2001), PrediSi (http://www.predisi.de/, last accessed July 23, 2013; Hiller et al. 2004), and SignalP 4.0 (http://www.cbs.dtu.dk/services/SignalP/, last accessed July 23, 2013; Petersen et al. 2011) softwares, while TMpred (http://www.ch.embnet.org/software/TMPRED_form.html, last accessed July 23, 2013; Hofmann and Stoffel 1993), Phobius (http://phobius.sbc.su.se/, last accessed July 23, 2013; Käll et al. 2004), InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/, last accessed July 23, 2013; Zdobnov and Apweiler 2001), Prodiv-TMHMM (http://topcons.cbr.su.se/, last accessed July 23, 2013; Bernsel et al. 2009), and Rhythm (http://proteinformatics.charite.de/rhythm/index.php?site=references, last accessed July 23, 2013) were used to localize putative transmembrane helices (TM-helices). Atome 2 (http://atome.cbs.cnrs.fr/AT2/meta.html, last accessed July 23, 2013; Pons and Labesse 2009), I-Tasser (http://zhanglab.ccmb.med.umich.edu/I-TASSER/, last accessed July 23, 2013; Zhang 2008), and HHpred (http://toolkit.tuebingen.mpg.de/hhpred, last accessed July 23, 2013; Söding et al. 2005) were used to find similarities with known proteins and to find clues on the possible functions of the mtORFs. Alignments of the putative novel mitochondrial proteins were performed with PSI-COFFEE (http://tcoffee.crg.cat/apps/tcoffee/do:psicoffee, last accessed July 23, 2013; Di Tommaso et al. 2011).

Mitochondrial novel ORFs recently found in Cnidaria were included in the function analysis for comparison: two putatively active proteins, DNA polymerase beta (PolB) (*Alatina moseri*: Cnidaria, Cubozoa, Alatinidae) (Smith et al. 2011) and DNA mismatch repair protein (mtMutS) (*Incrustatus comauensis*: Cnidaria, Anthozoa, Clavulariidae) (McFadden and van Ofwegen 2013), and ORF-314 (*Pelagia noctiluca*:

Table 1
Sequences Used in the Analyses

Species	mt Genome	Accession Number	ORF
Mollusca, Bivalvia			
Musculista senhousia	F	GU001953	Mse-FORF, Mse-ORF-B
		KC243365–75	Mse-FORF
		KC243354–64	Mse-ORF-B
	M	GU001952	Mse-ORF-B
		KC243376–87	Mse-ORF-B
Mytilus californianus	F	AY515227	Mca-FORF
	M	AF188284	Mca-MORF1, Mca-MORF2
Mytilus edulis	F	AY350784	Med-FORF
	M	AY823623	Med-MORF
Mytilus galloprovincialis	F	AY497292	Mga-FORF
	M	HM027630	Mga-MORF
Mytilus trossulus	F	GU936625	Mtr-FORF
	M	AF188282	Mtr-MORF
Ruditapes philippinarum	F	AB065375	Rph-FORF
		KC243324–31	Rph-FORF
	M	AB065374	Rph-MORF
		KC243347–53	Rph-MORF
Venustaconcha ellipsiformis	F	FJ809753	Vel-FORF
	M	FJ809752	Vel-MORF
Paphia euglypta		GU269271	Peu-ORF
Cnidaria			
Pelagia noctiluca		JN700949	Pno-ORF314
Alatina moseri		YP_005353032.1	Amo-PolB
Incrustatus comauensis		AFU34533.1	Ico-mtMutS

NOTE.—Mitochondrial genome type is specified only for ascertained DUI species. ORF column is the name given to the amino acid sequence.

Cnidaria, Scyphozoa, Discomedusae) (Kayal et al. 2011) (supplementary material S2, Supplementary Material online). Last accession to databases was in September 2012. p-distances of amino acid sequences of each novel ORFs were calculated using the bootstrap method with MEGA5 (Tamura et al. 2011). Percentage of amino acid difference of novel proteins and of all mtDNA-encoded protein genes were calcultated with MEGA5 (as in Breton et al. 2011a). For the *Myt. edulis* species complex (i.e., *Myt. edulis*, *Myt. Galloprovincialis*, and *Myt. trossulus*), pairwise sequence difference was first calculated for each gene and the results were then exported to Microsoft Excel for calculations of means and standard deviations (SDs).

Results

Novel Mitochondrial Open Reading Frames in Bivalves

The obtained *M. senhousia* LUR (FLUR of 11 females, 4,518–4,643 bp; MLUR of 12 males, 2,812–2,854 bp) and FUR2 (11 females, 542–543 bp) sequences were deposited in GenBank (FLUR accession nos.: KC243354–64; MLUR accession nos.: KC243376–87; FUR2 accession nos.: KC243365–75). The fORF, found in FUR2 on the heavy strand (as all standard coding genes) (fig. 1), is conserved in all samples (supplementary fig. S1, Supplementary Material online): its start and stop codons are always ATC and TAA, respectively, and its length is always 366 bp (121 aa). For nucleotidic p-distance see table 2. Another ORF, ORF-B, has been identified in MLUR and FLUR in the middle of Subunits B, on the reverse strand (fig. 1). In all males, ORF-B is always 318 bp long and its start and stop codons are ATG and TAA, respectively (supplementary fig. S2, Supplementary Material online). In females, Subunit B is duplicated (fig. 1) and ORF-B is not conserved as in males. The start codon is always ATG, and the stop codons can be TAA or TAG. Subunit B can contain one complete ORF-B (342–408 bp; supplementary fig. S2, Supplementary Material online) or two overlapping ORFs, together forming an ORF-B, due to a deletion of one T in a five-T string which breaks the frame. Two females showed only the version

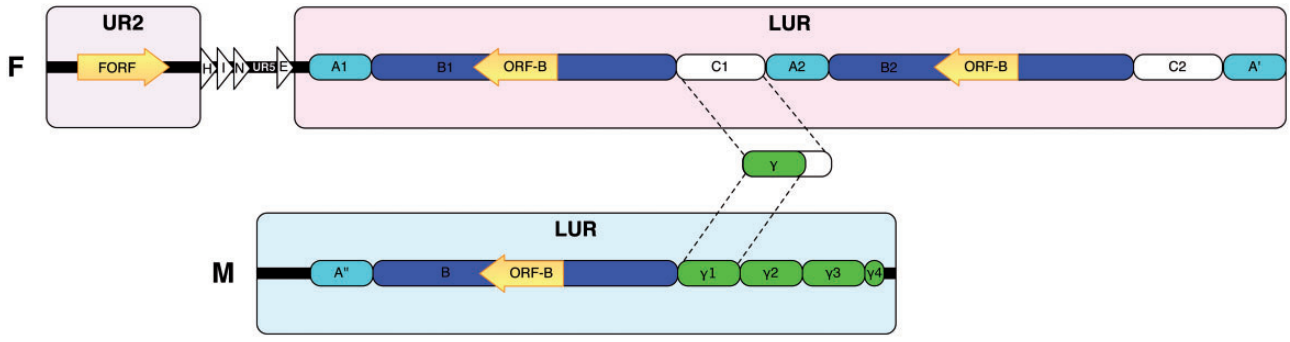

Fig. 1.—Largest Unassigned Regions (LURs). Schematic structure of female (F) and male (M) LURs of *Musculista senhousia*. Triangles indicate tRNAs.

Table 2

p-Distance (p-D) and Standard Error Values of Novel Mitochondrial ORFs in DUI Bivalves

Species	ORF	Nucleotide		Translation		*N*
		p-D	SE	p-D	SE	
Musculista senhousia	fORF	0.019	0.004	0.035	0.010	11
	Male ORF-B	0.004	0.002	0.008	0.004	12
	Female ORF-B[a]	0.024	0.005	0.056	0.012	8
	Overall ORF-B[b]	0.030	0.006	0.063	0.014	20
Mytilus californianus	fORF	0.005	0.003	0.014	0.008	4
	mORF1	0.015	0.009	0.031	0.021	4
	mORF2	0.011	0.007	0.033	0.022	4
Mytilus edulis	fORF	0.013	0.002	0.026	0.006	134
	mORF	0.017	0.004	0.039	0.012	25
Mytilus galloprovincialis	fORF	0.024	0.004	0.048	0.009	16
	mORF	0.029	0.008	0.062	0.021	47
	mORF (edulis-like)[c]	0.023	0.007	0.042	0.017	14
Mytilus trossulus	fORF	0.007	0.002	0.014	0.005	8
	mORF	0.025	0.007	0.046	0.016	9
Ruditapes philippinarum	fORF	0.009	0.003	0.011	0.006	8
	mORF	0.004	0.002	0.000	0.000	7
Venustaconcha ellipsiformis	fORF	0.000	0.000	0.000	0.000	3

Note.—Number of ORF sequences used for each species is dependant on the number of available and suitable sequences on GenBank. p-distances of *Myt. edulis*, *Myt. galloprovincialis*, and *Myt. trossulus* mORFs were calculated only on the last part of the ORF immediately following the poly-A sequence (see text for details). *N* = number of sequences used.

[a]Only complete female ORF-B were considered.

[b]Male ORF-B and complete female ORF-B were considered.

[c]mORF sequences matching *Myt. edulis* mORF.

with the two overlapping ORFs, never showing the complete ORF-B sequence (supplementary fig. S2, Supplementary Material online).

mt LURs of four *Mytilus* species (GenBank accession nos. in supplementary table S1, Supplementary Material online) were searched for the presence of the novel lineage-specific ORF described in Breton et al. (2011b): only the longest f- and mORFs were considered, as the shortest ones are often parts of them. A total of 201 *Mytilus* sequences containing complete ORFs were found (downloaded in September 2012): 197 fORFs and 17 mORFs. Many mORFs were found showing frame-disrupting mutations (supplementary table S1, Supplementary Material online). These alterations were more common in the first part of the expected mORF in *Myt. edulis*, *Myt. galloprovincialis*, and *Myt. trossulus*, before and inside a long poly-A sequence (from 17 to 48 nucleotides), while the last part is usually conserved in comparison to the ORFs described in Breton et al. (2011b). p-distances of *Myt. edulis*, *Myt. galloprovincialis*, and *Myt. trossulus* mORFs, because of alignment issues, were calculated only on the part of the ORF following the poly-A sequence. As indicated by the p-distance analysis (table 2), *Mytilus* spp. fORFs are less variable than mORFs. In *R. philippinarum* the situation is the opposite, as mORF is more conserved than fORF. For *V. ellipsiformis* only three fORF sequences were available, but they show a remarkable conservation. An ORF was also found in the LUR of the venerid *P. euglypta*.

Putative Novel Proteins from Bivalve Mitochondrial ORFs

Table 1 and supplementary material S2, Supplementary Material online, show sequences of the analyzed novel ORFs. A global alignment including all the analyzed amino acid sequences was not possible due to their divergence (supplementary fig. S3, Supplementary Material online), but groups with some similarities were found. Mse-ORF-B translation has practically the same amino acid sequence in the two genomes (supplementary fig. S4, Supplementary Material online). Mytilid FORFs are largely similar among each other (fig. 2*A*), most of all those of *Myt. edulis* complex (Med-, Mga-, and Mtr-FORFs) (supplementary fig. S5, Supplementary Material online). With the only exception of Mca-MORFs, *Mytilus* MORFs are also highly similar (fig. 2*B*; supplementary fig. S6, Supplementary Material online), and show a characteristic string of lysines (poly-K region) of variable length (8–12 aa; translation of a poly-A nucleotide sequence), absent from MORFs of other species and from FORFs. Downstream the poly-K region, *Mytilus* MORFs show a high similarity among each other, whereas in their N-terminus they are quite variable (supplementary fig. S6, Supplementary Material online). Although *Mytilus* FORFs and MORFs appear different between each other (see for example *Myt. edulis*, fig. 3*A*), Rph-FORF and MORF show several shared domains (fig. 3*B*), and also Vel-FORF and MORF have a big domain in their N-terminal showing similarity (supplementary fig. S7, Supplementary Material online).

Shared domains among the novel putative proteins are boxed in figure 4. Amino acid p-distances are reported in table 2. A common feature of all ORFs amino acid sequences (with the exception of *R. philippinarum* MORF and *V. ellipsiformis* FORF) is their major p-distance value in respect to their own nucleotidic sequences: this indicates that non-synonymous mutations are more common than synonymous mutations. The variability of FORFs and MORFs was confirmed by the amino acid sequence difference analysis of all mtDNA-encoded protein genes (fig. 5). Our findings, together with previous studies (Breton et al. 2009, 2011a), showed that lineage-specific mitochondrial proteins are among the fastest evolving proteins coded by the mtDNA of the analyzed species.

A SP was found in the N-terminus of all FORFs (table 3). Among the TM-helices, the N-terminal helix coincides with the SP sequence (table 3). Besides this helix, one more TM-helix supported by at least two programs was found in Mga-FORF, in Mtr-FORF, and in Rph-FORF (table 3). A sound SP was not always found in MORFs, even if some softwares point to the same SP sequence with a low score (table 3). Also in this case, the N-terminal TM-helices coincide with the SP sequence. Other probable TM-helices detected by at least two of the softwares were found in Mse-ORF-B, Med-MORF, and in Rph-MORF (table 3).

Novel Mitochondrial ORFs: Function Prediction

Atome 2, I-Tasser, and HHpred found domains similar, in structure or ligands, to known proteins, in both FORFs (tables 4, 5 and supplementary tables S2–S8, Supplementary Material online) and MORFs (tables 4, 5 and supplementary tables S9–S16, Supplementary Material online). FORF highest probability hits include proteins involved in nucleic acid binding and transcription (e.g., helicase/hydrolase, transcription factors), in some cases with specific aspects of nucleic acid processing, like RNA modification (e.g., Med-FORF and Vel-FORF), and methylation (e.g., Mtr-FORF). Other hits are proteins with a membrane association, for example involved in transport across membrane, in cell adhesion, but also receptors, most of all involved in hormone signalling. Many proteins point to a role in immune response, for example in cytokine release for immune system activation (e.g., Mca-FORF).

MORF hits with the highest probability include membrane-associated proteins with a role in nucleic acid binding and transcription, mainly related to signalling for cell differentiation and development (e.g., embryonic development). Some ORFs appear to be involved in DNA recombination and repair, in transposition regulation, and DNA integration of foreign elements (e.g., Mca-MORF1 and Rph-MORF). Moreover, several hits are proteins that regulate cytoskeleton formation and dynamics, from cell polarity regulation to cell proliferation. Other hits point to a role in ubiquitination and apoptosis with high probability (e.g., Mca-MORF1, Med-MORF, and Rph-MORF). Finally, many of the proteins have a role in immune response, for example in cytokine release (e.g., Mca-MORF2 and Med-MORF).

We found similar hits in Peu-ORF and Pno-ORF314 (tables 4 and 5 and supplementary tables S17 and S18, Supplementary Material online), connected with nucleic acid binding and transcription, with membrane association (Pno-ORF314), with signalling for cell differentiation during embryogenesis, with foreign elements (mobile genetic element and viral proteins), and with immune response regulation (Pno-ORF314).

All the hits come from different animal and plant proteins, from both unicellular and pluricellular organisms. The position of the most represented functional domains is reported in figure 5 (see also table 1 for acronyms). On the whole, with the only exception of Mtr-MORF and Vel-MORF, every analyzed protein showed hits referred to viral proteins (table 5 and fig. 5). In some cases (Mse-FORF, Mca-FORF, Mse-ORF-B, and Rph-MORF) the similarity with viral proteins was confirmed by all the three softwares used, in other cases (Mtr-FORF, Mca-MORF, Med-MORF, and Mga-MORF) by two of the softwares, and for the remaining proteins (Med-FORF, Mga-FORF, Rph-FORF, and Vel-FORF) by one program. Moreover, the same first four hits found by HHpred are present in all the novel putative proteins analyzed (supplementary table S19, Supplementary Material online), except for Amo-PolB, which showed complete homology with base-excision repair DNA

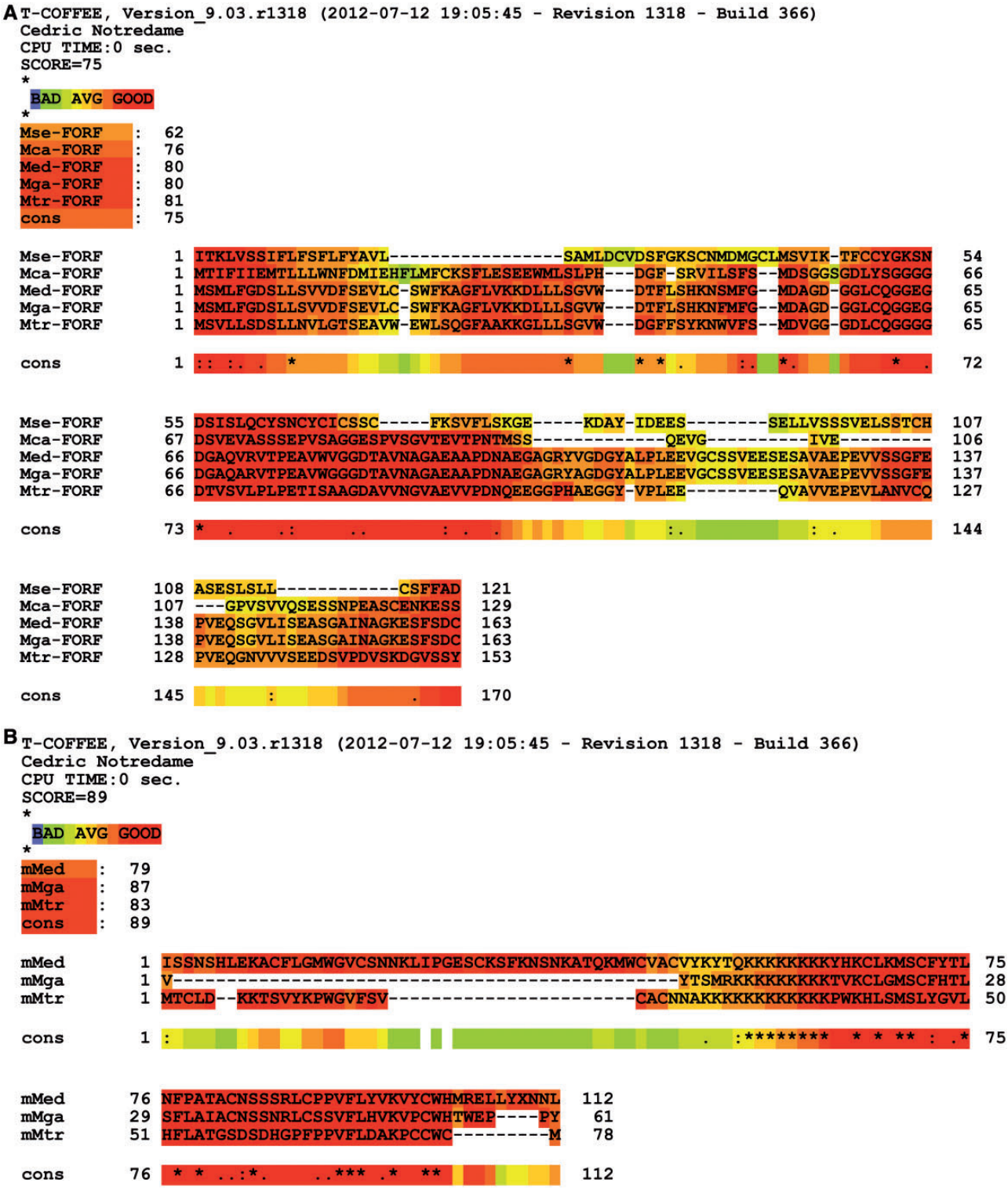

FIG. 2.—(*A*) PSI-Coffee alignment of FORFs of family Mytilidae (accession nos.: GU001953, AY515227, AY350784, AY497292, GU936625); (*B*) PSI-Coffee alignment of MORFs of *Mytilus* species (accession nos.: AY823623, HM027630, AF188282).

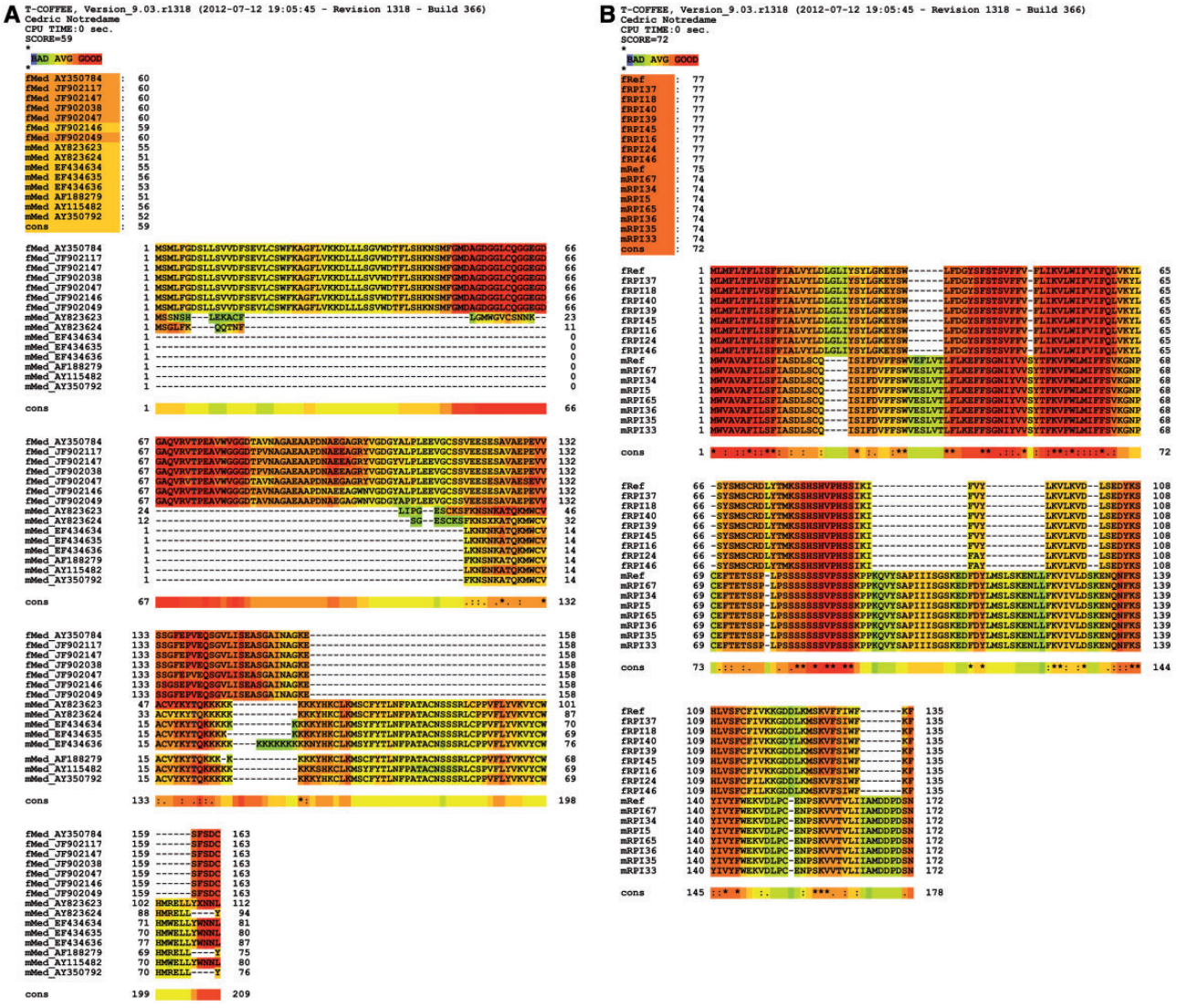

Fig. 3.—(*A*) PSI-Coffee alignment of *Mytilus edulis* FORF and MORF (accession nos. of sequences containing the ORF are reported in the figure); (*B*) PSI-Coffee alignment of *Ruditapes philippinarum* FORF (accession nos. of entire FLURs: KC243324–31) and MORF (accession nos. of entire MUR21 sequences: KC243347–53).

polymerases, mainly polymerase beta (HHpred probability: 100.0), and Ico-mtMutS, which showed a complete homology with a DNA mismatch repair protein (HHpred probability: 100.0), in both cases with hits from many organisms.

Discussion

Novel ORFs Characterization

As mentioned, mt genomes of bivalve species with DUI have novel lineage-specific ORFs of unknown origin and function. Generally, homologous proteins, or their fragments, have similar structure because structures diverge much more slowly than their sequences (Chothia and Lesk 1986). Depending on the degree of divergence between them, homologous proteins may also maintain similar cellular function, ligands, protein interactions partners, or enzymatic mechanisms (Todd et al. 2001). Because bivalve novel ORFs do not have known homologous (i.e., they are ORFans; Fischer and Eisenberg 1999), we performed multiple analyses of their structure, in order to infer the function. These ORFs are found in extragenic regions, often inside the LUR. Except for *M. senhousia* ORF-B, that is found in both mt genomes (in the middle of LUR Subunit B), the other analyzed ORFs are lineage-specific. ORF-B nucleotide sequence is extremely conserved between the two mt genomes (supplementary fig. S2, Supplementary Material online), but considering that in some *M. senhousia* females the complete ORF-B is absent, ORF-B might not be functional in females.

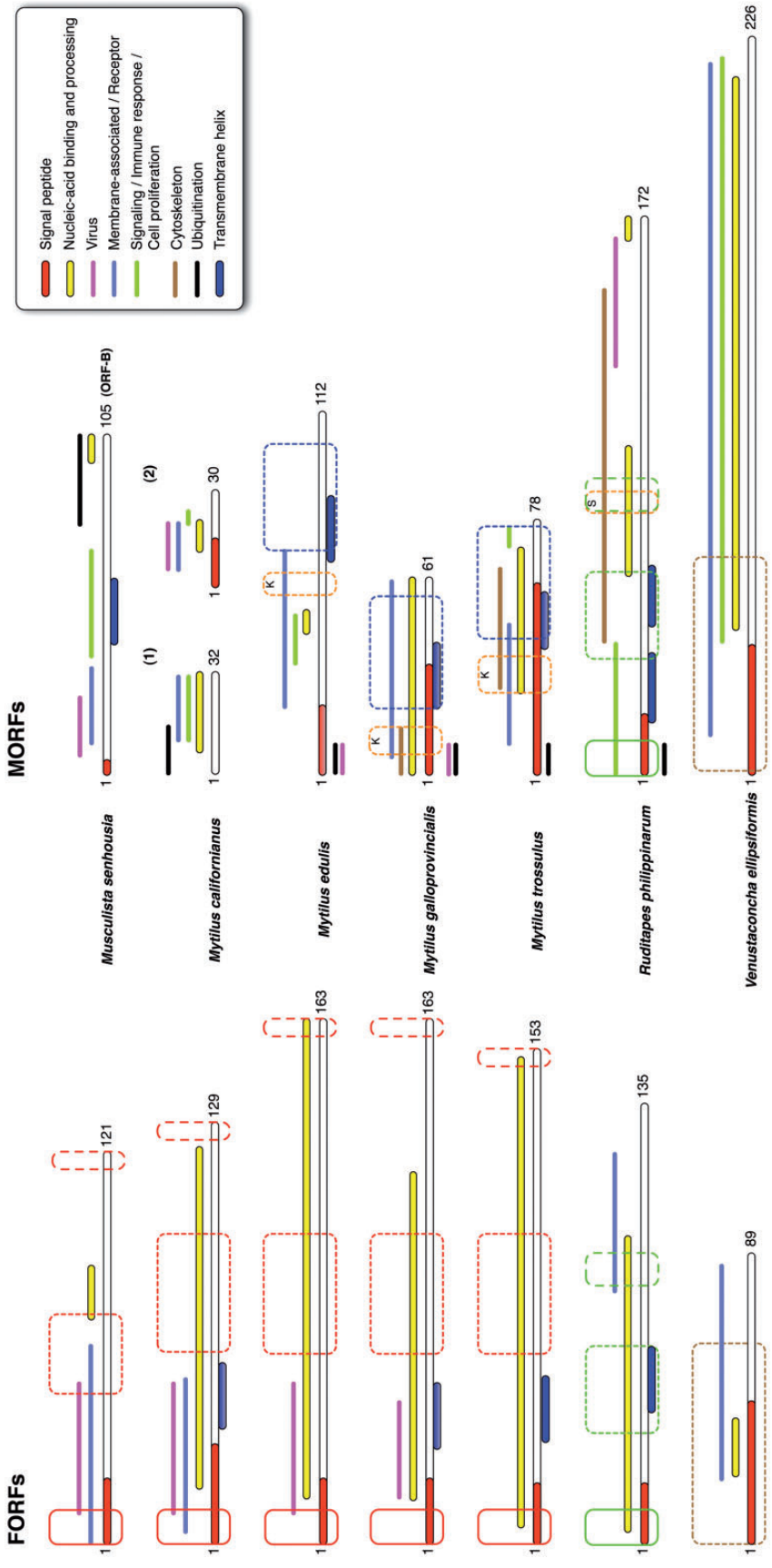

Fig. 4.—Functional domains in FORFs and MORFs (position in the amino acid sequence as identified by HHpred). Sequences with similarities are boxed in the same color and with the same type of line; red: similarities among FORFs; blue: similarities among MORFs; orange: K = poly-K region; S = poly-S region (see also PSI-Coffee alignments, figs. 2 and 3 and supplementary tables S2–S18, Supplementary Material online). Numbers indicate sequence length.

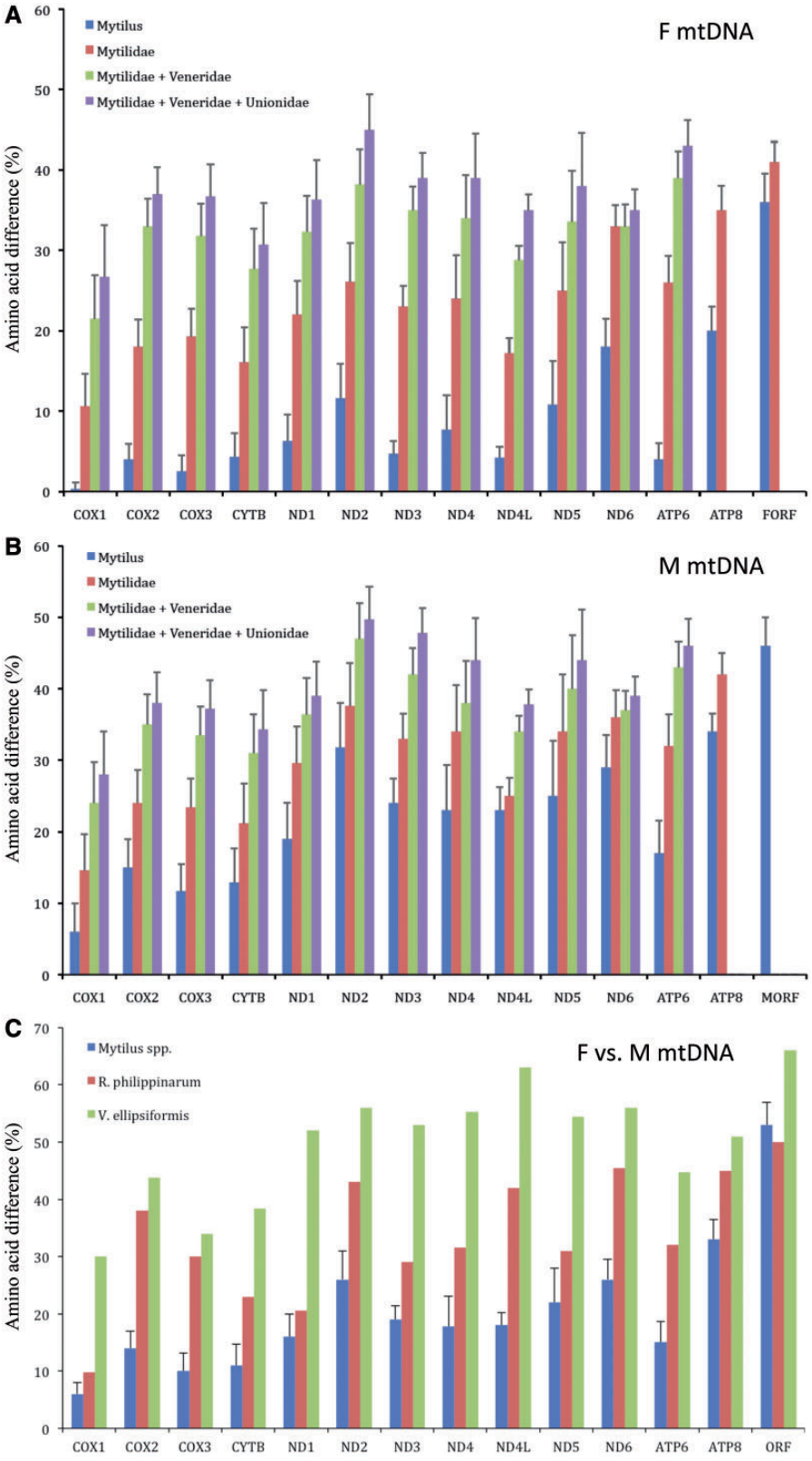

Fig. 5.—Percentage of amino acid difference of novel proteins and of all mtDNA-encoded protein genes. Amino acid divergence (% amino acid difference) was calcultated with MEGA5 for each mt protein coding gene among: (*A*) F mt genomes [for (i) *Mytilus* spp.; (ii) Mytilidae, i.e., *Mytilus* spp.

(continued)

Lineage-specific mitochondrial ORFs were found in all the analyzed DUI species (table 1; supplementary material S2, Supplementary Material online). In *Mytilus* male genomes, the last part of the mORFs, after the poly-A region, is the most conserved (fig. 4). A number of mORFs found in sequences annotated as *Myt. galloprovincialis* are identical to *Myt. edulis* mORF, and probably derive from hybridization that is extremely common inside the *Myt. edulis* complex: these "edulis-like" mORFs are more conserved than *Myt. galloprovincialis* own mORF and fORF, but are more diverse than *Myt. edulis* own mORF, from which they seem to derive (table 2). Nonetheless, *Myt. edulis* complex mORFs could be the same element, considering the extreme conservation of most of their sequence. Instead, *M. californianus* has two largely overlapping putative mORFs that do not contain a poly-A sequence like the other three species and are completely diverse from them. This is not surprising given the high divergence between *Myt. edulis* complex and *M. californianus* mitochondrial genomes (Zouros 2012).

Putative TM-helices were not found in all the analyzed proteins. In some cases the same region was identified as SP (table 3): being SP a peptide chain of hydrophobic amino acids, it can be difficult for softwares to discern it from a TM-helix (Käll et al. 2004). A clue in favour of a membrane association of MORFs comes from the poly-lysine (Med-, Mga-, and Mtr-MORF) and poly-serine (Rph-MORF) regions. Poly-lysine motif is required for membrane lipid binding (Bouaouina et al. 2012), and poly-serine domains characterize proteins anchored to bacterial outer membrane (Howard et al. 2004). Being mitochondria derived from alpha-proteobacteria (Andersson et al. 1998), we can hypothesize a similar membrane association in these organelles. Interestingly, the first four hits found with HHpred are the same for both FORFs and MORFs of DUI bivalves (supplementary table S19, Supplementary Material online), and for Peu-ORF and Pno-ORF314. Two of these hits are involved in the anchor to cell membrane/surface (LPXTG-motif cell wall anchor domain and outer membrane insertion C-terminal signal); the other two are typical of proteins involved in transcription (X-X-X-Leu-X-X-Gly heptad repeats) and in post-transcriptional processes (pentatricopeptide repeats, PPR). The detected motifs are not long enough to claim a functional homology, but their involvement in membrane binding and in transcription is sustained also by other hits (see tables 4 and 5; supplementary tables S2–S16, Supplementary Material online).

The existence of Vel-FORF and MORF was shown by western blot analysis (Breton et al. 2009), and Vel-FORF was shown to be present in mitochondria and in the nuclear membrane (Breton et al. 2011a). Likely, these novel mitochondrial proteins have a role in different cellular compartments, thus including domains that allow them to interact with several substrates such as membranes, cytoskeleton, and nucleic acids. It is important to investigate the existence of ORF translation products in other DUI species. We are performing these kind of analyses and first data confirm the existence of Rph-MORF protein (Milani et al. in preparation). Furthermore, increasing the number of analyzed DUI species and sequences may help in explaining the evolutionary dynamics that led to the highest similarity found between FORF and MORF of some species (i.e., Rph-FORF/-MORF and Vel-FORF/-MORF) in comparison to other species (i.e., *Myt. edulis* complex) (see alignments and fig. 4).

The similarity region between an ORF and a known protein sometimes includes a large part of the protein, even with high probability (see for example Vel-MORF), in other cases, as said before, it is found in short amino acid sequences. In such cases we are confident we retrieved sound similarities, because the same homolog proteins from very distant taxa, from both unicellular and pluricellular organisms, are present among the hits (supplementary tables S2–S16, Supplementary Material online).

Overall, the analyzed ORFs show many common functions (see supplementary tables S2–S16, Supplementary Material online), but, when we consider only hits with the highest scores, FORFs are more similar among each other than with MORFs, and vice-versa (tables 4 and 5). FORFs appear to be involved in transcription regulation and in immune response, also linked to cell adhesion, migration, and proliferation. MORFs appear to have a main role in cytoskeleton organization (cell differentiation during embryonic development), but also capable, as FORFs, of nucleic acid binding and transcription regulation. FORFs and MORFs appear to share a role as

Fig. 5.—Continued

and *Musculista senhousia*; (iii) Mytilidae + the venerid *Ruditapes philippinarum*; and (iv) Mytilidae + the venerid *R. philippinarum* + the unionoid *Venustaconcha ellipsiformis*], (*B*) M mt genomes [for (i) *Mytilus* spp.; (ii) Mytilidae, i.e., *Mytilus* spp. and *M. senhousia*; (iii) Mytilidae + the venerid *R. philippinarum*; and (iv) Mytilidae + the venerid *R. philippinarum* + the unionoid *V. ellipsiformis*], and (*C*) between F and M mt genomes [for (i) *Mytilus* spp.; (ii) *R. philippinarum*; and (iii) *V. ellipsiformis*]. For the *Mytilus edulis* species complex (i.e., *Myt. edulis*, *Myt. galloprovincialis*, and *Myt. trossulus*), pairwise sequence difference was first calculated for each gene and the results were then exported to Microsoft Excel for calculations of means and SDs. For both *R. philippinarum* and *V. ellipsiformis* only, one whole F mtDNA and one whole M mtDNA are present in database and no error can be calculated. Omitted comparisons are due to the impossibility to obtain a good alignment. Note: F mtDNA = female mitochondrial genome; M mtDNA = male mitochondrial genome. *Mytilus* spp. = *Myt. edulis* species complex. Accession nos. mitochondrial genomes (F-type and M-type mtDNA, respectively): *Myt. edulis* NC_006161 and AY823623; *Myt. galloprovincialis* NC_006886 and AY363687; *Myt. trossulus* DQ198231 and DQ198225; *M. senhousia* GU001953 and GU001954; *R. philippinarum* AB065375.1 and AB065374.1; *V. ellipsiformis* FJ809753 and FJ809752.

Table 3
Signal Peptide and Transmembrane-Helix Prediction in the Novel Putative Proteins

Signal Peptide							
FORF	**Mse**	**Mca**	**Med**	**Mga**	**Mtr**	**Rph**	**Vel**
Software							
Phobius	1–20	—	1–20*	1–20*	1–18*	1–18	—
InterProScan	1–20	1–31	—	—	1–18	1–18	1–44
PrediSi	1–28	—	1–20*	1–20*	1–18*	1–18	1–44*
SignalP 4.0	1–20	1–20*	1–20*	1–20*	1–18*	—	1–44*
MORF	**Mse^{ORFB}**	**Mca**	**Med**	**Mga**	**Mtr**	**Rph**	**Vel**
Software							
Phobius	—	- (1–13)	—	1–34	—	1–18	—
InterProScan	1–5	- (1–18)	—	—	—	1–18	1–40
PrediSi	—	- (1–16)*	1–22*	1–34*	1–59	1–17	1–40
SignalP 4.0	1–6*	- (1–14)	1–21*	1–34*	1–59*	1–18	1–40*
Transmembrane Helices							
FORF	**Mse**	**Mca**	**Med**	**Mga**	**Mtr**	**Rph**	**Vel**
Software							
TMpred	4–23	3–25	8–29	8–29	**31–52**	1–18/**40–59**	21–42
Phobius	—	—	—	—	—	7–27/**39–62**	21–42
InterProScan	—	—	—	—	—	5–23/**42–62**	21–41
Prodiv-TMHMM	5–27	5–25/**35–55**	9–29	5–25/**28–48**	**26–47**	3–23/**39–59**	21–42
Rhythm	6–23	4–23	—	**18–37**	**33–50**	5–27/**40–62**	21–42
MORF	**Mse^{ORFB}**	**Mca**	**Med**	**Mga**	**Mtr**	**Rph**	**Vel**
Software							
TMpred	**40–61**	- (-)	**69–96****	19–35	41–57**	1–23/**16–38/46–64**	21–39
Phobius	—	- (-)	—	—	38–56	**42–62**	20–38
InterProScan	—	- (-)	—	20–38	38–56	5–27/**41–61**	21–41
Prodiv-TMHMM	**41–61**	- (-)	**65–86**	21–41	38–59	3–23/**44–64**	—
Rhythm	—	- (-)	—	17–34	41–57	**20–37/46–64**	21–38

NOTE.—Signal peptide: Only signal peptides statistically supported (Phobius posterior label probability > 0.5; PrediSi score > 0.5; SignalP score > D-cutoff 0.5; significance test not provided by InterProScan) or found at least by two softwares are shown; *Significance < 0.5; (*n*) = Mca-MORF2 results. Transmembrane helices: Only transmembrane helices considered significant (TMpred score > 500; Phobius posterior label probability > 0.5; significance test not provided by the other softwares) or found by at least two softwares are shown; **TMpred score < 500; values in bold indicate helices not overlapping with the predicted signal peptide; (*n*) = Mca-MORF2 results.

signalling molecules, more specifically involved in hormone signalling and immune response regulation. Interestingly, some MORFs show similarity with DNA replication, recombination, and repair proteins (see for example the transposition regulation and DNA binding-integration hits of Mca-MORF1). Moreover, hits of ubiquitination and apoptosis regulation proteins are found in almost all ORFs (supplementary tables S2–S16, Supplementary Material online).

Are Novel Mitochondrial ORFans of Viral Origin?

The sequences analyzed in this article do not show homologies with any known mitochondrial protein, therefore they unlikely originated from recent duplication events, as instead happened for *nad2* in *Crassostrea* (Wu et al. 2012), for *cox2* in *R. philippinarum* F-mtDNA (Okazaki M and Ueshima R, unpublished data), and in *M. senhousia* M-mtDNA (Passamonti et al. 2011). Another origin should be taken into account for these proteins and the observed hits to viral proteins provide a possible working hypothesis: bivalve ORFs could have arisen from different events of insertion, thus showing a narrow distribution similar to other ORFans (Yu and Stoltzfus 2012).

The analyzed ORFs show a higher amino acid substitution rate than the typical mitochondrial coding genes (fig. 5). Lineage-specific genes evolve at a faster rate than broadly distributed genes, in both bacteria and eukaryotes (Daubin and Ochman 2004a, 2004b; Yu and Stoltzfus 2012). One reason could be that lineage-specific genes participate more in lineage-specific adaptation, therefore evolving faster

Table 4
Function Analysis of Novel Mitochondrial ORFs

Mse-FORF	Mca-FORF
Hormone receptor/Cell adhesion, migration, proliferation/Immune response	**Transport across membrane/Receptor/Immune response**
Atome2 (highest probability):	Atome2 (highest probability):
Chemokine (13), highest score 75.17	Unique short US2 glycoprotein, score 82.16
Human tissue factor, score 70.69	Killer cell immunoglobulin-like receptor 2DL1 (2), highest score 75.39
Eotaxin (2), score 67.89 and 62.51	Pertussis toxin subunit 5, score 55.44
Erythrocyte binding antigen 175, score 54.15	Putative ABC type-2 transporter, score 54.92
I-Tasser (confirmation):	I-Tasser (confirmation):
Cell division protein kinase 9/Protein Tat, *Z*-score 0.79	Receptor-type adenylate cyclase (2), TM-score > 0.5
RhoGAP protein, *Z*-score 0.90	HHpred (confirmation):
Glypican-1, *Z*-score 0.62	RAB6-interacting protein 2 (2), highest probability 62.09, aa 39–99
Small-inducible cytokine A13, *Z*-score 0.91	Integral membrane protein, probability 48.70, aa 10–38
Erythrocyte binding antigen 175, *Z*-score 0.63	TonB Periplasmic protein TonB, probability 45.44, aa 44–75
HHpred (confirmation):	NIPSNAP, probability 35.39, aa 12–34
SARS receptor-binding domain-like, probability 54.78, aa 31–61	Membrane protein, probability 31.83, aa 54–66
Small inducible cytokine A1 precursor, probability 27.01, aa 5–91	Membrane or secreted protein, probability 30.95, aa 3–51
Protein binding/transport	Transport protein Sec24A (2), probability 30.88, aa 3–51
I-Tasser (highest probability):	Membrane protein containing DUF1112, probability 28.02, aa 14–66
Exportin-5, *Z*-score 0.75	**Cell adhesion and migration/Hormone receptor**
Cullin-5, *Z*-score 0.69	Atome2 (highest probability):
Nucleoporin NUP170, *Z*-score 0.84	Fibronectin, score 70.61
BRO1 protein, TM-score > 0.5	PfEMP1 variant 2 of strain MC, score 58.80
GTP-binding nuclear protein Ran, TM-score > 0.5	Human tissue factor (2), highest score 52.67
Nucleic acid binding	**Helicase activity/Replication/Immune response**
I-Tasser (highest probability):	I-Tasser (highest probability):
Telomeric repeat-binding factor (2), TM-score > 0.5	Antiviral helicase SKI2, *Z*-score 0.68
ATP-dependent RNA helicase (2), TM-score > 0.5	Proliferating cell nuclear antigen PcnA, *Z*-score 0.85
Membrane association	Infectivity protein G3P, *Z*-score 0.62
HHpred (highest probability):	Cyclophilin-like domain, *Z*-score 0.59
More than 40 hits, highest probability 75.84, aa 1–21	HHpred (confirmation):
Transcription factor translocator	SKI2/RNA helicase, probability 50.91, aa 101–123
HHpred (highest probability):	Peptidyl-prolyl isomerase G/cyclophilin G, probability 38.59, aa 17–69
Glucocorticoid receptor-like (10), highest probability 64.44, aa 68–85	**Cytoskeleton/Cytokine release/Immune system activation**
	HHpred (highest probability):
	Keratin (9 hits), highest probability 76.37, aa 25–95
	Transcription regulator
	HHpred (highest probability):
	Sterol regulatory element binding protein (2), highest probability 71.01, aa 17–66
	CG17964-PH, isoform H, probability 28.03, aa 54–122

Med-FORF	Mga-FORF
DNA binding and replication	**DNA binding and replication**
Atome2 (highest probability):	Atome2 (highest probability):
Uncharacterized protein AF_1548, score 85.31	Uncharacterized protein AF_1548, score 82.33
Exotoxin A, score 74.29	Exotoxin A, score 72.53
Minichromosome maintenance protein, score 63.74	Minichromosome maintenance protein, score 62.01
I-Tasser (highest probability):	I-Tasser (highest probability):
Minichromosome maintenance protein (3), highest *Z*-score 1.64, TM-score > 0.5	Minichromosome maintenance protein (2), highest *Z*-score 1.64, TM-score > 0.5
ATPase involved in replication control (3), highest *Z*-score 1.31, TM-score > 0.5	ATPase involved in replication control (3), highest *Z*-score 1.21, TM-score > 0.5
P97 (Cell division cycle), TM-score > 0.5	HHpred (confirmation):
HHpred (confirmation):	Zinc fingers (2), highest probability 36.31, aa 13–32, 36–42
Zinc fingers (2), highest probability 35.41, aa 13–32 and 36–42	NPH4/transcription factor, probability 30.20, aa 31–114

(continued)

Table 4 Continued

Med-FORF	Mga-FORF
Development/Growth hormone receptor/Cell adhesion Atome2: Nicotinamidase, score 59.05 Human tissue factor (2 hits), highest score 52.26 Fibronectin, score 51.93 **Lyase/Hydrolase activity** HHpred (highest probability): Cyanase C-terminal domain (2), highest probability 54.52, aa 5–16 **Immune response/RNA binding and processing** HHpred (highest probability): Cyclophilin/Peptidylprolyl isomerase (13), highest probability 46.01, aa 59–163 **Cell adhesion/Lipid metabolism** HHpred (highest probability): GYF domain (2 hits), highest probability 39.73, aa 16–28 Malonyl-CoA decarboxylase (4), highest probability 36.39, aa 14–29	**Development/Growth hormone receptor/Cell adhesion** Atome2: Human tissue factor, score 56.15 Fibronectin, score 52.65 Nicotinamidase, score 44.75 Tudor domain-containing protein 5 (Germ line integrity), score 43.50 **Lyase/Hydrolase activity** HHpred (highest probability): Cyanase C-terminal domain (2), highest probability 55.41, aa 5–16 **Lipid metabolism/Cell adhesion** HHpred (highest probability): Malonyl-CoA decarboxylase (6), highest probability 39.89, aa 14–29 GYF domain (3 hits), highest probability 39.13, aa 16–28

Mtr-FORF	Rph-FORF
Ligase activity Atome2 (highest probability): D-alanine—poly(phosphoribitol) ligase subunit 1 (3), highest score 80.77 **Lipid metabolism** Atome2 (highest probability): Acetyl-coenzyme A synthetase (3), highest score 79.06 **Receptor/Membrane-associated protein/Immune response** Atome2: Unique short US2 glycoprotein, score 65.52 Interleukin 18 binding protein/Cytokine, score 50.72 I-Tasser (confirmation): Gramicidin synthetase 1, *Z*-score 2.25 D-alanine—poly(phosphoribitol) ligase subunit 1, *Z*-score 2.12 **Cytoskeleton-associated protein** I-Tasser (highest probability): Kinesin-like protein Nod, TM-score > 0.5 Tubulin (3), TM-score > 0.5 Integrin alpha-X, TM-score 0.498 HHpred (confirmation): Actin-like ATPase domain (2), highest probability 45.12, aa 49–57 **Methylation (DNA, RNA, protein)** HHpred (highest probability): More then 20 hits (21), highest probability 61.81, aa 5–149 **Immune response/Viral infection cofactor** (large region) Cyclophilin, probability 26.70, aa 15–110	**Nuclear transport** Atome2 (highest probability): Nuclear transport factor 2 (2), highest score 85.83 NTF2-related export protein 1, score 79.17 I-Tasser (highest probability): Nuclear transport factor 2 (3), highest *Z*-score 0.72, TM-score > 0.5 p15 (Export of mRNAs through nuclear pore complexes) (2), TM-score > 0.5 Nuclear RNA export factor 2, TM-score > 0.5 mRNA transport regulator Mtr2, TM-score > 0.5 Rasputin (Similar to nuclear transport factor 2), TM-score > 0.5 HHpred (confirmation): DNA double-strand break repair transporter domain, probability 49.97, aa 77–88 **DNA replication/Transcription/Nucleic-acid binding** HHpred (highest probability): 20 hits Highest probability 86.90, with HemY family protein, aa 3–67 Zinc fingers, probability 86.88 Atome2 (confirmation): Polymerase PB2, score 56.13 Restriction endonuclease Hpy99I, score 52.92 Cyclin (3), highest score 52.40 DNA gyrase inhibitor YacG, score 50.57 **Transport across membrane/Amino-acid transporter** HHpred: About 30 hits, highest probability 86.47, aa 2–75 **Receptor site** HHpred: Neurotoxin type G, probability 63.95, aa 77–120 **Membrane-associated protein/Immune response** HHpred: Macoilin/transmembrane protein 57 (2), probability 50.56, aa 1–113 LysM domain, probability 33.34, aa 115–123 Atome2 (confirmation): HLA class II histocompatibility antigen, score 47.74

(continued)

Table 4 Continued

Vel-FORF
Nuclear proteins/Nuclear transport/RNA processing
Atome2 (highest probability):
Poly(A) polymerase, score 84.27
Ran GTPase-activating protein 1, score 60.97
Chimera of Histone H2B.1 and Histone H2A.*Z*, score 49.66
I-Tasser (confirmation):
VP1/mRNA-capping machine (2), highest *Z*-score 0.82
Poly(A) polymerase (2), highest *Z*-score 0.70
ATP-dependent DNA helicase RecG-related protein, *Z*-score 0.71
DNA binding/Transcription
Atome2 (highest probability):
Bifunctional protein GlmU, score 58.95
Serine/threonine-protein phosphatase (2), highest score 58.24
SAGA-associated factor 73, 21.79
HHpred (highest probability):
ComGC (2), highest probability 94.43, aa 2–38
CG13581-PA transcription factor, probability 39.77, aa 77–89
Membrane-associated proteins
Atome2 (highest probability):
Bactericidal permeability-increasing protein, score 53.65
Photosystem II reaction center protein I, score 31.82
HHpred (confirmation):
More than 10 hits in the N-terminus of the sequence
Hormone receptor/Transcription
I-Tasser (highest probability):
Progesterone receptor ligand-binding domain, TM-score > 0.5
Androgen receptor ligand-binding domain, TM-score > 0.5
AncCR, TM-score > 0.5
Mineralocorticoid receptor (nuclear receptor), TM-score > 0.5
Immune system/Transport across membrane
HHpred:
C-type LECtin family member (clec-35) (7), highest probability 78.84, aa 19–86

Mca-MORF1	Mca-MORF2
Transposition regulation/DNA binding and integration/Transcription	**Protein folding**
Atome2 (highest probability):	Atome2 (highest probability):
Transposase (3), highest score 76.46	Huwentoxin-II, score 79.45
Protein RDM1/RNA-directed DNA methylation, score 65.30	Alanine racemes, score 76.95
Modification methylase TaqI, score 55.99	Heat shock 70 kDa protein 8/Chaperone (2), highest score 55.37
Nuclear factor NF-kappa-B p100 subunit, score 55.42	BAG-family molecular chaperone regulator-1, score 47.88
Replication termination protein, score 54.13	**Cytokine/Immune response/Cell proliferation/Embryonic development**
DNA-binding protein RAP1, score 50.90	Atome2 (highest probability):
I-Tasser (confirmation):	Interleukin-6 receptor subunit beta, score 71.60
C25G10.02, chromosome I (Hydrolase/DNA duplexes separation), *Z*-scores > 1	Interleukin-1 beta, score 40.82
Rad50 (Hydrolase/DNA-double strand break repair), *Z*-scores > 1	Erythropoietin receptor, score 54.98
Replication factor c small subunit, TM-scores > 0.5	Tumor necrosis factor ligand superfamily member 13, score 50.90
O-sialoglycoprotein endopeptidase/protein kinase (Hydrolase), TM-scores > 0.5	Natural killer cell activating receptor, score 46.35
HHpred (highest probability):	Myeloid antimicrobial peptide 27, score 41.93
"Winged helix" DNA-binding domain (2), highest probability 88.51, aa 7–25	Tumor necrosis factor receptor associated protein 2, score 41.37
	T-cell immunoglobulin and mucin domain-containing protein 4, score 39.37
	I-Tasser (confirmation):
	Tumor protein P73 (cell cycle control), *Z*-score > 1

(continued)

Table 4 Continued

Mca-MORF1	Mca-MORF2
C2H2 and C2HC zinc fingers (3), highest probability 80.42, aa 16–32	**Membrane association**
Transcription factor E2F-4, winged-helix (2), highest probability 60.60, aa 7–16	Atome2 (highest probability):
Hormone signaling	Rieske protein, score 71.25
I-Tasser (highest probability):	NADH-cytochrome b5 reductase 3, score 70.00
Parathyroid hormone (4), *Z*-score > 1	ATP synthase subunit alpha, score 39.91
HHpred (confirmation):	**DNA replication, recombination, and repair**
Kazal-type inhibitors/growth factor receptor (9), highest probability 68.60, aa 13–20	HHpred (highest probability):
Apoptosis	Methylated DNA-protein cysteine methyltransferase (24), highest probability 80.02, aa 13–19
I-Tasser (highest probability):	I-Tasser (confirmation):
Apoptosis regulator BCL-2 (4), TM-scores > 0.5	DNA topoisomerase I, TM-score > 0.5
Apoptosis regulator BAK, TM-score > 0.5	**Receptor/Signaling (Immune response)**
Signaling/Regulation of cytoskeleton formation/Cell proliferation	HHpred (highest probability):
HHpred (highest probability):	XII secretory phospholipase A2 precursor, probability 76.62, aa 18–24
GTPase-activator protein (47), highest probability 87.22	Toxin_33/Waglerin family (acetylcholine receptor), probability 70.71, aa 11–20
Ubiquitination	Immunoglobulin domain (12), highest probability 62.63, aa 5–21
HHpred (highest probability):	Tumor necrosis factor receptor superfamily member 17 (2), highest probability 60.38, aa 16–25
UBA-like (4), highest probability 72.94, aa 1–15	
Membrane association	
HHpred (highest probability):	
Tim10-like/Mitochondrial translocase (2), highest probability 68.38, aa 19–28	
Atome2, confirmation:	
Photosystem I reaction center subunit IX, score 43.87	

Med-MORF	Mga-MORF
Membrane association	**Cytoskeleton dynamics/Cell proliferation and differentiation/Hormone signaling**
Atome2 (highest probability):	Atome2 (highest probability):
Alcohol dehydrogenase 4/Oxidoreductase, score 77.67	FGFR1 oncogene partner, score 88.04
I-Tasser (highest probability):	HIV-1 envelope protein chimera/Chemokine receptor, score 59.63
AP-2 complex subunit beta-2, *Z*-score 0.64	Filamin-binding LIM protein 1, score 55.61
Ubiquitination	Sprouty-related, EVH1 domain-containing protein 1, score 34.00
Atome2 (highest probability):	Vasodilator-stimulated phosphoprotein, score 30.93
UPF0147 protein Ta0600/Ubiquitin-conjugating enzyme E2, 72.27	Protein enabled homolog, score 26.03
Cytokine/Receptor/Immune response	Proliferation-associated protein 2G4, score 25.29
I-Tasser (highest probability):	I-Tasser (confirmation):
Complement C5A anaphylatoxin, *Z*-score 0.61	Gamma filamin (2), highest *Z*-score 0.72
Glutathione S-transferase omega-2, *Z*-score 0.58	HHpred (highest probability):
Discoidin domain receptor 2, *Z*-score 0.74	Actin, probability 87.91, aa 1–16
Receptor protein-tyrosine kinase erbB-3, *Z*-score 0.55	EPS8/epidermal growth factor receptor kinase substrate 8-like protein 1, probability 71.11, aa 4–15
Interleukin-13, *Z*-score 0.63	**Immune response**
Coagulogen, *Z*-score 0.56	I-Tasser (highest probability):
Atome2 (confirmation):	Glutathione S-transferase (5 hits), TM-scores > 0.5
Interleukin-12 subunit alpha, score 51.43	**Transcription factor/Nucleic-acid binding/Differentiation and development**
Tumor necrosis factor alpha-induced protein 3, score 44.65	HHpred (highest probability):
HHpred (highest probability):	Helix-loop-helix (bHLH) protein, Human Nulp1 (2), highest probability 87.71, aa 3–16
Glutathione transferase domain/Thioredoxin (3), highest probability 67.32, aa 33–49	
CG33975-PA/Glucocorticoid induced gene 1, probability 64.83, aa 20–51	

(continued)

Table 4 Continued

Med-MORF	Mga-MORF
Nuclear Hormone Receptor family, probability 61.43, aa 28–69	PEP-CTERM putative exosortase interaction domain, probability 59.70, aa 1–10
Transcription	Sp1 transcription factor, probability 57.36, aa 5–61
HHpred (highest probability):	Josephin domain containing 3, probability 50.13, aa 6–15
Zinc finger protein 395 and 704, highest probability 57.58, aa 43–51	Kruppel-like factor (Growth-factor pathways), probability 48.23, aa 54–61
SLC2A4 regulator, 52.80, aa 43–51	**Signal transduction/Cell proliferation**
Glycoprotein/Membrane association/Cell–cell connection	HHpred:
HHpred:	Smoothened homolog (2), highest probability 81.43, aa 4–21
Protocadherin (13), highest probability 59.36, aa 53–61 (poli-K region, aa 55–62)	**Membrane-associated protein/Hormone receptor**
	HHpred:
	Extracellular solute-binding protein (2), highest probability 74.34, aa 5–60
	EEV glycoprotein, probability 69.62, aa 7–36
	Lipoprotein, probability 56.70, aa 5–39
	FIG1, Factor-induced gene 1 protein (Mating/Pheromone-regulated membrane protein) (2), highest probability 51.63, aa 28–47
	Glycoprotein/Membrane association/Cell–cell connection
	HHpred:
	Protocadherin (26), highest probability 75.96, aa 7–14 (poli-K region, aa 7–15)

Mtr-MORF	Rph-MORF
Growth hormone receptor/Cell adhesion, migration, proliferation during embryonic development	**Ubiquitination factors**
Atome2 (highest probability):	Atome2 (highest probability):
Human tissue factor (2), highest score 90.71	26S proteasome regulatory subunit rpn10, score 78.22
Skeletal dihydropyridine receptor, score 61.37	HHpred (confirmation):
Angiostatin, score 47.55	Zinc ion binding, ubiquitin interaction motif-containing protein (2), highest probability 59.68, aa 72–95
Fibronectin, score 46.00	NEDD8 ultimate buster-1/Ubiquitin-like protein, probability 41.84, aa 73–96
Membrane-binding proteins	**Membrane association**
Atome2 (highest probability):	Atome2 (highest probability):
Complexin (2), highest score 52.74	L-aspartate dehydrogenase/Oxidoreductase, score 71.39
HHpred (confirmation):	Transient receptor potential cation channel subfamily V member 1, score 59.26
N-acetylglucosaminyl-phosphatidylinositol de-n-acetylase, probability 76.89, aa 9–38	Unique short US2 glycoprotein, score 34.59
Membrane protein, probability 75.99, aa 26–47	**Transcription**
Cell growth and differentiation/signaling	I-Tasser (highest probability):
I-Tasser (highest probability):	Archaeal transcriptional regulator TrmB, *Z*-score 1.04
T-lymphoma invasion and metastasis-inducing protein 2, *Z*-score 0.75	Atome2 (confirmation):
C3, *Z*-score 0.64	Tumor suppressor p53-binding protein 1, score 57.01
KEX1(DELTA)P, Prohormone-processing serine carboxypeptidase, *Z*-score 0.74	HHpred (confirmation):
Cell differentiation	Restricted Tev Movement 2 (hormone receptor), probability 41.60, aa 61–94
HHpred:	Forkhead-associated phosphopeptide binding domain 1 isoform 19, probability 31.88, aa 68–101
Gametogenetin binding protein 2, probability 71.72, aa 3–40	Exonuclease, probability 30.93, aa 69–99
Microtubule association	**Immune resistance**
HHpred:	HHpred (highest probability):
Kinectin 1 microtubule-dependent transport, probability 68.69, aa 26–63	CRISPR-associated DEAD/DEAH-box helicase Csf4, probability 71.11, aa 144–165
Nucleic-acid binding/Transcription factor/DNA repair ATPase	

(continued)

Table 4 Continued

Mtr-MORF	Rph-MORF
HHpred (highest probability): Helix-loop-helix (bHLH) protein; Human Nulp1 (2), highest probability 95.13, aa 23–37 Telomeric telomer cycle, DNA-binding, protein binding, probability 68.11, aa 51–64 PHD FINGER domain, probability 62.04, aa 25–63 DNA double-strand break repair ATPase Rad50, probability 61.51, aa 42–70 **Signaling** HHpred: Cysteine alpha-hairpin motif, probability 65.87, aa 70–77 **Glycoprotein/Membrane association/Cell–cell connection** HHpred: Protocadherin, highest probability 74.29, aa 25–37 (poli-K region, aa 25–37)	**Cytoskeleton organization/Cell proliferation, migration, differentiation/Immune response** HHpred (highest probability): Structural maintenance of chromosomes (3), highest probability 63.66, aa 65–140 Translation proteins SH3-like domain, 58.57, aa 61–75 RAD50 (4), highest probability 35.41, aa 163–172 Subunit of MRX complex with Mre11p and Xrs2p, probability 29.87, aa 163–172 Gelsolin (6), highest probability 46.96, aa 40–146 Villin (6), highest probability 36.65, aa 40–146 C15A11.5/Collagen family member, probability 42.97, aa 1–41 CG14217-PB, isoform B (Serine threonine kinase), probability 42.82, aa 69–91 Mitochondrial tumor suppressor 1 isoform 5, probability 38.92, aa 65–101 EGF/Laminin, probability 32.22, aa 64–99 Keratin (2), highest probability 30.68, aa 63–109 Segment polarity protein Dishevelled (Development), probability 29.40, aa 66–94 CG12047-PC, isoform C (Centrosome/spindle organization), probability 28.75, aa 65–78 Atome2 (confirmation): Thymosin beta-4, score 36.83 Adseverin, score 36.03 I-Tasser (confirmation): Proliferating cellular nuclear antigen 1, *Z*-score 1.03 Guanine nucleotide-binding protein G(q) subunit alpha, *Z*-score 0.61 Chimera of Gelsolin domain 1 and C-Terminal domain of thymosin Beta-4, *Z*-score 0.74

Vel-MORF	Mse-ORF-B
Protein folding Atome2 (highest probability): Chaperone protein ClpB (2), highest score 89.09 **Actin cytoskeleton and cell polarity regulator/Cell differentiation and adhesion/Cell cycle** Atome2 (highest probability): Myosin-7 (2), highest score 82.66 Rho-associated protein kinase 1, score 65.91 Tropomyosin alpha-1 chain, score 53.16 DNA topoisomerase 4 subunit A, score 53.11 Cell division protein ZapB, score 52.81 I-Tasser (highest probability): ATP-dependent helicase/nuclease subunit A, *Z*-score 1.19 YIIU, *Z*-score 0.61 Spectrin (4), highest *Z*-scores 1.19 Myosin-5A, *Z*-score 1.19 Cdc42-interacting protein 4, *Z*-score 0.63 Desmoplakin, TM-score 0.47 HHpred (confirmation): Keratin (6) (cytokine release/immune system), highest probability 94.31, aa 81–171	**Cytoskeleton organization/Cell adhesion, migration, proliferation/Immune response** Atome2 (highest probability): Myomesin-1, score 90.17 Fibronectin, score 66.05 Fibrinogen-binding protein, score 32.61 **Hormone receptor** Atome2 (highest probability): Human tissue factor (hormone signaling/cell adhesion) (2), highest score 82.66 HHpred (confirmation): F11G11.10/Collagen family member, probability 41.10, aa 36–69 Alpha-actinin, probability 38.91, aa 70–85 TyrPK_CSF1-R (Cytokine/Immune response), probability 31.97, aa 95–102 Fibrinogen-binding protein/cell adhesion complex (3), highest probability 30.60, aa 82–93 PDGF Platelet-derived and vascular endothelial growth factors, probability 21.10, aa 13–28 **Membrane association** Atome2 (highest probability):

(continued)

Table 4 Continued

Vel-MORF	Mse-ORF-B
Laminin (5) (cytokine release/immune system), highest probability 93.43, aa 41–218	Unique short US2 glycoprotein, score 77.87
Membrane protein/ Receptor/Immune response	I-Tasser (highest probability):
Atome2 (highest hits):	mRNA export factor Mex67 (Associated to nuclear pores), *Z*-score 0.90
C-Jun-amino-terminal kinase-interacting protein 4 Isoform 4 (Sperm surface protein), score 73.95	**Signaling**
HHpred (highest probability):	I-Tasser (highest probability):
More than 20 hits of antigens, all probabilities higher than 90, aa 12–218	Sensor protein (3 hits), TM-scores > 0.5
Nuclear pore complex proteins, 15 hits, all probabilities higher than 90, aa 41–220	**Nucleic acid binding/Immune response**
I-Tasser (confirmation):	HHpred (highest probability):
Sensor protein (3), TM-scores > 0.5	Recombination-activating protein 2 (2), highest probability 79.31, aa 9–33
Methyl-accepting chemotaxis transducer (MCPs), TM-score > 0.5	Nucleic acid-binding proteins (4), highest probability 74.26, aa 94–105
Invasin IPAD, TM-score > 0.5	I-Tasser (confirmation):
Cell invasion protein SIPD, TM-score > 0.5	Transcription intermediary factor 1-alpha, *Z*-score 0.66
Pathogenicity island 1 effector protein, TM-score > 0.5	DNA polymerase sliding clamp C, *Z*-score 0.66
Translocator protein bid, TM-score > 0.5	
Toll-like receptor 5b and variable lymphocyte receptor B.61 chimeric protein, TM-score > 0.5	
Transcription factor/Nucleic-acid binding and transport	
HHpred:	
Basic leucine zipper (bZIP) transcription factor (2), highest probability 92.14, aa 44–171	
Nucleotide binding, probability 91.76, aa 90–213	
mRNA localization machinery, probability 90.81, aa 50–171	

Peu-ORF	Pno-ORF314
Cell differentiation during embryogenesis/Hormone receptor	**Nucleic-acid binding and transcription**
Atome2 (highest probability):	Atome2 (highest probability):
Cytoplasmic FMR1-interacting protein 1, score 61.36	Small protein B, score 82.73
Tumor necrosis factor alpha/Cytokine, score 54.49	ATP-dependent RNA helicase SUPV3L1, mitochondrial, score 69.03
Atrial natriuretic peptide receptor A, score 52.42	DNA topoisomerase 4 subunit A, score 50.41
Mesoderm development candidate 2, score 44.10	I-Tasser (highest probability):
I-Tasser (confirmation):	Anti-sigma F factor (Prokaryote gene expression regulation) (6), highest *Z*-score 0.68
Mesoderm development candidate 2, *Z*-score 0.73	Transcriptional regulator LRPA (2), highest *Z*-score 0.64
Cytoplasmic FMR1-interacting protein 1, *Z*-score 0.92	Conserved domain protein/Transcriptional regulator, score 0.57
HHpred (confirmation):	Bromodomain and PHD finger-containing protein 3; SPOIIAA, score 0.69
FnI-like domain (Cell adhesion/migration during embryonic development) (4), highest probability 62.25, aa 52–64	HHpred:
Jun-like transcription factor/Mitogen-activated protein kinases (Cellular responses to cytokines/Cell proliferation/differentiation), probability 50.47, 2–26	Histone-fold (2), highest probability 58.93, aa 62–77
Resistin/Cytokine (2), highest probability 46.20, aa 49–63	CCAAT-BOX DNA binding protein subunit B, probability 50.87, aa 64–77
DNA replication	**Cell differentiation during embryogenesis**
I-Tasser (highest probability):	Atome2 (highest probability):
Proliferating cell nuclear antigen, *Z*-score 0.81	Mesoderm development candidate 2, score 79.76
DNA polymerase processivity factor, *Z*-score 0.69	**Membrane association**
Poly [ADP-ribose] polymerase 15, *Z*-score 0.63	I-Tasser (highest probability):
Flap structure-specific endonuclease (DNA repair/replication), *Z*-score 0.70	Sulfate transporter, TM-score 0.608
HHpred (confirmation):	**Viral protein**
Proliferating cell nuclear antigen, probability 42.24, aa 6–22	HHpred (highest probability):
	8 hits, highest probability 82.37, aa 3–59

(continued)

Table 4 Continued

Peu-ORF	Pno-ORF314
Immune resistance	I-Tasser (highest probability):
HHpred (highest probability):	Capsid protein P27 (2), highest *Z*-score 0.92
CRISPR-associated DxTHG motif protein, probability 75.05, aa 4–17	**Protein folding**
Nucleic-acid binding/Transcriptional regulator	HHpred (highest probability):
HHpred (highest probability):	LDLR chaperone BOCA, probability 77.86, aa 2–52
More than 40 hits, highest probability 60.91, aa 34–66	**Immune response**
	HHpred:
	Immunoglobulin domain, probability 45.90, aa 79–103

NOTE.—Hits with the highest probability are reported for each of the three programs together with eventual confirmation of the same biological process from the other two softwares. Norm. *Z*-score > 1 = good alignment; TM-score > 0.5 = similar fold with query (Zhang 2008; Xu and Zhang 2010); (*n*) = number of the same hit (protein), when more than one. See also supplementary tables S2–S16, Supplementary Material online.

(Cai and Petrov 2010). Similarly, the lineage-specific novel mtORFs may experience such a kind of evolutionary pressure, maybe for features related to sexual differentiation.

A large amount of pathways toward new gene origin through the domestication of parasitic genome sequences has been documented (Kaessmann 2010). In addition to their infectious properties, which enable them to spread horizontally between individuals and across species, many viruses can also become part of the genetic material of their host, a process that is called endogenization: endogenous viruses have integrated into the germ line of their host, allowing for vertical transmission and fixation in the host population (Boeke and Stoye 1997; Belshaw et al. 2004; Feschotte and Gilbert 2012). Viruses are able to integrate both in eukaryote and prokaryote genomes: for example, ORFans present in bacterial genomes are hypothesized to have been acquired through horizontal transfer from viruses (Daubin and Ochman 2004a, 2004b). Quite remarkably, the initiator protein DnaC in bacteria and the mitochondrial DNA replication and transcription apparatus have been recently documented to have a viral origin (Forterre 2010 and references therein). In the light of what reported above about endogenization in prokaryotes, a viral origin of novel mitochondrial genes is not unconceivable.

Novel ORFs were recently found also in the linear mitochondrial genome of Medusozoa. Using the same approach as for bivalve novel ORFs, we found a complete homology of Amo-PolB with the polymerase beta of several organisms and of Ico-mtMutS with a DNA mismatch repair protein (thus confirming the results obtained by Smith et al. 2011 and McFadden and van Ofwegen 2013, respectively). In both cases, the function of the novel mitochondrial proteins is supported. Instead, even if the product of ORF314 was proposed to act in concert with PolB in the maintenance of chromosome ends, it did not show a sound similarity with any other protein in database (Kayal et al. 2011). Interestingly, we found that it shares many predicted functions with the novel mitochondrial ORFs of bivalves (supplementary table S18, Supplementary Material online). In fact, almost all the analyzed bivalve ORFs, together with Pno-ORF314, show hits pointing to immune response and viral proteins (tables 4 and 5). Viruses can manipulate the host cell molecular machinery to counteract antiviral defences and to control the expression of their own genes, moreover viral sequences can be co-opted for host cell functions (Feschotte and Gilbert 2012), contributing to host genome evolution. For example, a viral gene has been co-opted to serve an important function in the physiology of mammals: syncytin is the envelope gene of a human endogenous defective retrovirus and is important in human placental morphogenesis and probably in the immune tolerance of the developing embryo (Mi et al. 2000). Interestingly, recent data attest that some genes involved in mammal placental development derive from domestication of multiple retrovirus-derived genes (Nakagawa et al. 2013). Similarly, we think that virus-derived novel mitochondrial proteins may have acquired new functions in the host. All the analyzed ORFs show an involvement in transcription regulation, like many virus-derived sequences that have been incorporated into the regulatory system of mammalian genes (Britten and Davidson 1969; Feschotte 2008; Cohen et al. 2009).

Role in Immune Response and Apoptosis

Microbial invasion generally causes an immune reaction (Galluzzi et al. 2008). Mitochondria play a central role in primary host defence mechanisms against viral infections, and a number of viral proteins interact with mitochondria to regulate cellular responses (Ohta and Nishiyama 2011). Once viruses infect their hosts, they activate signalling pathways leading to the production of specific molecules (i.e., chemokines and cytokines) (Bryant and Fitzgerald 2009; Takeuchi and Akira 2009), and viruses have developed strategies to evade host immune responses: because signalling from recognition receptors converges in mitochondria, it is plausible that viruses would target mitochondrial processes to evade immune responses (Ohta and Nishiyama 2011). A clue in favor of an interaction between novel mitochondrial ORFs and immune system comes from the many hits pointing to

Table 5
Hits to Viral Proteins Found in Novel Mitochondrial ORFs

DUI sp.	Hits	Position
FORF		
Mse	Protein Tat [Atome2; score 54.94] (**Nuclear transcriptional activator of viral gene expression/Cell division**)	n.a.
	Protein Tat [I-Tasser; norm. *Z*-score 0.79]	n.a.
	Protein Tat [HHpred; probability 25.94]	62–73
	SARS receptor-binding domain-like [HHpred; 54.78]	31–61
	Hepatitis E virus ORF-2 (**Capsid protein/Pro-apoptotic gene expression activation/Host–cell cytoplasm**) [HHpred; 23.74]	61–69
	Fijivirus P9-2 protein (**Unknown function**) [HHpred; probability 23.19]	8–50
Mca	Unique short US2 glycoprotein (**Viral protein/Transport across membrane/Immune recognition masking**) [Atome2; score 82.16]	n.a.
	Pre-neck appendage protein (Bacteriophage) (5 hits) [Atome2; score 57.87–51.81]	n.a.
	Antiviral helicase SKI2 [I-Tasser; norm. *Z*-score 0.68]	n.a.
	Infectivity protein G3P (**Viral protein**) [I-Tasser; norm. *Z*-score 0.62]	n.a.
	Cyclophilin-like domain (**Viral infection cofactor/RNA and protein processing**) [I-Tasser; norm. *Z*-score 0.59]	n.a.
	Phage small terminase subunit (**DNA binding/Endonuclease activity/Viral capsid assembly**) [HHpred; probability 44.52]	8–45
Med	Retrovirus capsid dimerization domain-like (2) [HHpred; probability 35.34, 29.28]	14–43
Mga	Retrovirus capsid dimerization domain-like (2) [HHpred; probability 35.47, 30.09]	14–43
Mtr	Unique short US2 glycoprotein (**Viral protein/Transport across membrane/Immune recognition masking**) [Atome2; score 65.52]	n.a.
	Positive stranded ssRNA viruses [HHpred; probability 28.66]	16–54
Rph	Polymerase PB2 (**Polymerase; Viral RNA replication**) [Atome2; score 56.13]	n.a.
Vel	VP1, the protein that forms the mRNA-capping machine (**Viral protein**) (2) [I-Tasser; norm. *Z*-score 0.82, 0.70]	n.a.
	Fibritin (**Viral protein**) [I-Tasser; norm. *Z*-score 0.64]	n.a.
MORF		
Mca^{ORF1}	Early 35 kDa protein (**Apoptosis-preventing protein/Protease inhibitor/Response to the viral infection**) [Atome2; score 47.39]	n.a.
	Phosphatidylinositol 3-kinase regulatory subunit alpha (**Host–virus interaction/Signaling/Transferase**) [Atome2; score 44.26]	n.a.
	V-bcl-2 (**Viral protein/Apoptosis**) [I-Tasser; TM-score > 0.5]	n.a.
Mca^{ORF2}	Circulin A (**Cyclic peptide/Virus cytopathic effects and replication inhibitor**) [I-Tasser; norm. *Z*-score > 1]	n.a.
	First immunoglobulin (Ig) domain of nectin-3 (**Poliovirus receptor related protein 3/Cell adhesion**) [HHpred; probability 62.63]	12–21
	Coxsackie virus and adenovirus receptor (Glycoprotein A33; CTX-related type I transmembrane protein) [HHpred; probability 51.10]	5–21
	Coxsackie virus and adenovirus receptor (Car), domain 1 [*Homo sapiens*, TaxId: 9606] [HHpred; probability 49.70]	12–21
	Hepatitis A virus cellular receptor 1 [*Mus musculus*] [HHpred; probability 45.53]	12–25
Med	Replicase polyprotein 1ab (**Viral protein/RNA, DNA duplex-unwinding activities/ATPase/Deubiquitination**) [Atome2; score 58.58]	n.a.
	Macro domain of Non-structural protein 3 (**Viral protein/RNA binding protein**) [I-Tasser; norm. *Z*-score 0.70]	n.a.
Mga	HIV-1 envelope protein chimera (**Viral envelope glycoprotein/Chemokine receptor**) [Atome2; score 59.63]	n.a.
	Proliferation-associated protein 2G4 (**Viral Translation/Growth regulation/Androgen receptor/Transcriptional regulation**) [Atome2; score 25.29]	n.a.
	Viral protein [I-Tasser; norm. *Z*-score 0.72]	n.a.
Mtr	—	—
Rph	Unique short US2 glycoprotein (**Viral protein/Transport across membrane/Immune recognition masking**) [Atome2; score 34.59]	n.a.
	Viral protein/Signaling protein [I-Tasser; norm. *Z*-score 0.57]	n.a.
	CRISPR-associated DEAD/DEAH-box helicase Csf4 (**Phage genomic sequence insertion/Resistance against mobile genetic elements: viruses, transposable elements, conjugative plasmids**) [HHpred; probability 71.11]	144–165
	d.172.1 gp120 core (56502) SCOP seed sequence: d1g9mg_ (**Viral envelope receptor**) [HHpred; probability 34.78]	125–157
Vel	—	—

(continued)

Table 5 Continued

DUI sp.	Hits	Position
Mse^{ORFB}	Unique short US2 glycoprotein (**Viral protein/Transport across membrane/Immune recognition masking**) [Atome2; score 77.87]	n.a.
	Gag-Pol polyprotein (**Capsid protein/Host nucleus**) [Atome2; score 54.53]	n.a.
	Glycosyltransferase (Mannosyltransferase) (**Capsid viral protein/Transferase**) [I-Tasser; norm. *Z*-score 0.90]	n.a.
	VAC_I5L (dsDNA viruses, no RNA stage; Poxviridae) (**Membrane-associated protein**) [HHpred; probability 31.24]	6–24
Other sp.		
Peu	Terminase small subunit (**Viral protein**) [Atome2; score 56.08]	n.a.
	CAG38821 (**Viral protein**) [I-Tasser; norm. *Z*-score 0.77]	n.a.
	Terminase small subunit (**Viral protein**) [I-Tasser; norm. *Z*-score 0.84]	n.a.
	DNA polymerase processivity factor (**DNA binding/Transferase/Viral protein**) [I-Tasser; norm. *Z*-score 0.69]	n.a.
	CRISPR-associated DxTHG motif protein (**Phage genomic sequence insertion/Resistance against mobile genetic elements: viruses, transposable elements, conjugative plasmids**) [HHpred; probability 75.05]	4–17
Pno-ORF314	Capsid protein P27 (**Viral protein**) (2) [I-Tasser; norm. *Z*-score 0.92, 0.86]	n.a.
	Retrovirus capsid protein, N-terminal core domain (**Viral replication**) [HHpred; probability 82.37]	21–50
	RSV capsid protein {Rous sarcoma virus [TaxId: 11886]} [HHpred; probability 80.17]	21–59
	JSRV capsid, capsid protein P27; zinc-finger, metal-binding {Jaagsiekte sheep retrovirus} (**Viral protein**) [HHpred; probability 78.55]	21–59
	Capsid protein P27; retrovirus, N-terminal core domain {Mason-pfizer monkey virus} (**Viral protein**) [HHpred; probability 74.21]	21–59
	GAG polyprotein capsid protein P27; retrovirus, immature GAG{Rous sarcoma virus} (**Viral protein**) [HHpred; probability 48.94]	21–50
	Capsid protein P27; viral protein, retrovirus, GAG; 7.00 A {Mason-pfizer monkey virus} [HHpred; probability 44.98]	22–59
	Capsid protein; two independent domains helical bundles, virus/viral protein {Rous sarcoma virus} [HHpred; probability 43.53]	21–47
	Tat binding protein 1 (TBP-1)-interacting protein (TBPIP) (**Eukaryotic protein/Modulates the inhibitory action of human TBP-1 on HIV-Tat-mediated transactivation**) [HHpred; probability 38.93]	3–50

NOTE.—Norm. *Z*-score > 1 = good alignment; TM-score > 0.5 = similar fold with query (Zhang 2008; Xu and Zhang 2010); (*n*) = number of the same hit (protein); position: amino acid position in the query sequence; n.a. = non applicable.

receptors and signaling molecules involved in immune response (antigens and cytokines above all). Some of these hits are present in both FORFs (Mse-FORF, Mca-FORF, Mtr-FORF, Vel-FORF; supplementary tables S2, S3, S6, and S8, Supplementary Material online) and MORFs (Mca-MORF2, Med-MORF, Mga-MORF, Rph-MORF, Vel-MORF; supplementary tables S11–S13, S15, and S16, Supplementary Material online), as in other analyzed ORFs (Mse-ORF-B, Peu-ORF; supplementary tables S9 and S17, Supplementary Material online). In Vel-MORF, the homology region almost coincides with the whole sequence (table 4 and supplementary table S16, Supplementary Material online).

Proteins reported in literature as acting in bivalve immune response (Gestal et al. 2008, and references therein) have homology with the analyzed mitochondrial ORFs, as for example, tumor necrosis factors (see hits found in Vel-FORF, Mca-MORF2, Med-MORF, Peu-ORF; supplementary tables S8, S11, S12, and S17, Supplementary Material online), interleukins (a group of cytokines; hits found in Mtr-FORF, Mca-MORF2, Med-MORF; supplementary tables S6, S11, and S12, Supplementary Material online), transforming growth factor (Kruppel-like factor; hits found in Mse-FORF, Mga-MORF; supplementary tables S2 and S13, Supplementary Material online) and platelet-derived growth factor (hit found in Mse-ORF-B; supplementary table S9, Supplementary Material online). All the reported findings strongly support a link between these mitochondrial novel proteins and the immune response of bivalves.

Microbial invasion also has a role in apoptosis regulation (Galluzzi et al. 2008): viruses have acquired the capacity to control host cell apoptosis and inflammatory responses, thus evading immune reactions (Galluzzi et al. 2008). Mitochondria have a central role also in apoptosis and, for this reason, a number of viral proteins are targeted to mitochondria to regulate this mechanism. Interestingly, hits of structural analogues with apoptotic factors were found with high probability in Mca-MORF1 (apoptosis regulator BCL-2, four hits with TM-scores > 0.5, and apoptosis regulator BAK, TM-score > 0.5) (table 4). It is known that several viral polypeptides are homologues of host-derived apoptosis-regulatory proteins, such as members of the BCL-2 family (Galluzzi et al. 2008), some of which assemble on the mitochondrial membrane (Wei et al. 2001; Kuwana et al. 2002; Nutt et al. 2002).

Viral BCL-2 homologues (vBCL-2) do not show significant sequence similarity with their host counterparts, but exhibit high structural resemblance (White et al. 1991; Cuconati and White 2002). This seems exactly the case of Mca-MORF1, in which the similarity with both BCL-2 and BAK proteins was

detected in the structure, not in the sequence (supplementary table S10, Supplementary Material online). Interestingly, viral proteins with a three-dimensional folding similar to BCL-2 are glycoprotein always showing a transmembrane domain flanked by positively charged amino acids (typically lysines) and followed by an hydrophilic tail (Wang et al. 2002; Douglas et al. 2007; Kvansakul et al. 2007). This domain is required for both the mitochondrial outer membrane targeting and the anti-apoptotic function (Douglas et al. 2007; Kvansakul et al. 2007). Interestingly, all these characters are shared by *Mytilus* MORFs and Rph-MORF (the latter with serines instead of lysines). Moreover, in some FORFs (Med-FORF, Mga-FORF, and Peu-ORF; supplementary tables S4, S5, and S17, Supplementary Material online), N-terminal homeodomain (PHD)-like regions were found. Recently, several PHD-containing viral proteins have been identified to promote immune evasion by down-regulating proteins that govern immune recognition by functioning as E3 ubiquitin ligases (Coscoy and Ganem 2003). Other hits specifically related to E3 ubiquitin ligases were found (Mse-FORF, Rph-FORF, Vel-FORF, Mse-ORF-B, Mca-MORF2; supplementary tables S2, S7–S9, and S11, Supplementary Material online). For all above-mentioned, we propose that the novel ORFs here analyzed may have originated from viral elements with a function in immune response and apoptosis control.

Interaction with Cytoskeleton: Mitochondrial Segregation

MORFs, together with viral hits, show many hits related to cytoskeleton/cytoskeleton-binding proteins. For example, among viral hits we obtained capsid proteins and Transactivator of transcription (Tat) proteins, a regulatory protein that enhances the efficiency of viral transcription and alters microtubule dynamics, promoting proteasomal degradation and a mitochondrion-dependent apoptotic pathway (Chen et al. 2002; Aprea et al. 2006; Egelé et al. 2008). Envelope proteins generally induce a perinuclear clustering of mitochondria by altering cytoskeleton conformation, interacting for example with keratins and microtubules, thus promoting the aggregation of these organelles (Doorbar et al. 1991; Galluzzi et al. 2008). Taking into account that mitochondria appear to respond to some viral infection by migrating with viral tegument proteins (Ohta and Nishiyama 2011), we suggest that these novel ORFs might have a role in the aggregation and localization of mitochondria, producing the aggregated and dispersed patterns of distribution of spermatozoon mitochondria observed in early DUI embryos. Many other hits are connected with cytoskeleton, such as microtubule-binding proteins, actin-binding proteins, cytoskeleton proteins themselves, and proteins with a role in cytoskeleton organization (table 4). Interestingly, several endosymbiotic pathogens can use proteins expressed on their surface to ensure their survival and/or alter host processes. These surface proteins can cause cytoskeleton remodeling, as best demonstrated in *Listeria monocytogenes*: this endosymbiont induces actin to assemble on its surface, propelling it through the cytoplasm and allowing its transport between host cells, bypassing host defense mechanisms (Ireton and Cossart 1997, and references therein). It is possible that MORFs bind some cytoskeleton elements, and, if they were membrane-associated proteins, they could be responsible for spermatozoon mitochondria positioning in DUI embryos.

Targeting and Export of Mitochondrial Novel Proteins

It is well established that the nucleus regulates organelle gene expression through anterograde regulation (Woodson and Chory 2008 and references therein). On the other hand, several studies have recently demonstrated that signals from organelles regulate nuclear gene expression by retrograde signaling (Butow and Narayan 2004). It appears likely that, given the complex cross-talk between the nucleus and mitochondria, not only chemical messengers but also exported proteins may participate in transducing signals from mitochondrion to nucleus.

A deeply studied example is the retrograde signaling that characterizes plants with Cytoplasmic Male Sterility (CMS) (Abad et al. 1995; Fujii and Toriyama 2008; Nizampatnam et al. 2009). CMS is known to be associated with the expression of novel mitochondrial ORFs and the accumulation of these novel proteins at proper spatial or temporal development stages induces male sterility (Fujii and Toriyama 2008). Moreover, some of these proteins contain a hydrophobic N terminus, commonly found in membrane-bound proteins (Abad et al. 1995 and references therein) so that it was hypothesized that they are mitochondrial membrane-bound proteins that might lead to disruption of the mitochondrial membrane integrity in the anther tissues, leading to pollen death (Nizampatnam et al. 2009, and references therein). The possibility of binding membranes is a feature in common with the here studied novel bivalve ORFs. In fact, many hits of the novel bivalve mitochondrial ORFs we analyzed were identified as proteins with a function on the cytoplasmic side of mitochondrial outer membrane (table 4). For example, bivalve mitochondrial novel proteins may tag the surface of mitochondria: MORFs may have a role in the maintenance of sperm mitochondria aggregation in the first stages of development, possibly masking them from the degradation that normally affects mitochondria carried from sperm in species with the more usual maternal inheritance of mitochondria. This could be possible thanks to the features that novel ORFs share with anti-apoptotic factors. Maybe, a similar mechanism involving novel ORF integration in the mitochondrial genome of females makes FORFs responsible for the inheritance of F-type mitochondria in DUI species, but, in this case, no evident difference from a SMI mechanism for mitochondrial transmission could be seen.

The presence of mitochondrial proteins in diverse cellular extramitochondrial sites, such as endoplasmic reticulum and nucleus, supports the existence of specific export mechanisms by which certain proteins exit mitochondria (Soltys and Gupta 2000). Mitochondria are derived from bacteria from which they probably inherited protein exit pathways used to elude host defense mechanism before the endosymbiont became an essential organism. Some of these protein exit mechanisms might have been retained and/or modified in mitochondria, allowing certain mitochondrial proteins to have additional functions in other subcellular compartments (Soltys and Gupta 2000). For example, besides the export of mitochondrial ribosomes in the cytoplasm, some mitochondrially encoded proteins are present on the cell surface as histocompatibility antigens, and are therefore exported from mitochondria (Soltys and Gupta 2000, and references therein). These peptides derive from partial sequences of mitochondrial genes (e.g., N-terminus of NADH dehydrogenase subunit 1, in mouse and humans; internal region of ATPase 6, in rat) probably by proteolysis of parent molecules inside mitochondria or in the cytoplasm, before being transported to the cell surface (Soltys and Gupta 2000). More than one mechanism by which mitochondrial matrix macromolecules are exported may exist but the processes are not fully clear yet. For example, the presence versus the detachment by peptidase of part of the protein sequence (for example an N-terminal SP) was proposed to be the cause of the re-targeting of mitochondrial proteins, and the use of protein import machinery, the leakage from breaks in the mitochondrial membranes during fission and/or fusion, membrane fusion with other organelles (e.g., endoplasmic reticulum and nucleus), the existence of protein transporters, the autotransport through lipids (as observed for heat shock proteins), and vesicle-mediated export involving vesicle budding (as in gram-negative bacteria) are other proposed mechanisms (Soltys and Gupta 2000). In our case, given the presence of a SP in many of the analyzed ORFs, this N-terminal sequence may be used to target the proteins to sites outside mitochondria. It is possible that proteins with post-transcriptional cleavage of the SP remain attached at the mitochondrial outer membrane, whereas peptide complete with the SP may be targeted elsewhere in the cell.

The Origin of Mitochondrial Novel ORFs and Implications for DUI Evolution

As mentioned, many clues point to a viral origin of novel mitochondrial ORFs, even if the probability of the hits is sometimes low and the regions of similarity of short length (table 5). As in the case of ORFans, this can be due to the extreme limited sampling of viral sequences (Daubin and Ochman 2004a, 2004b; Lerat et al. 2005). Suttle (2005) estimated that the virus population size in the ocean alone is ~4×10^{30}, with a phage diversity of ~10^8 (Rohwer 2003). For this reason, a significant fraction of the ORFs without detectable viral homologs may have arisen from not yet sequenced or extinct viruses (Yin and Fischer 2006). Moreover, many ORFans may remain without viral homologs if they have experienced rapid evolution after the integration in the new genome, diverging to the extent that no homology to viral proteins is detectable (Charlebois et al. 2003; Domazet-Loso and Tautz 2003; Daubin and Ochman 2004a; Siew and Fischer 2004; Yin and Fischer 2006).

The co-option of such novel genes by viral hosts may have determined some evolutionary aspects of host life cycle, possibly involving mitochondria (Forterre 2006; Koonin 2006), and, as supposed for ORFans (Hendrix et al. 2000; Juhala et al. 2000), bivalve mtORFs might now be involved in key cellular functions. The study of novel mitochondrial proteins expression during the bivalve life cycle could help in understanding their function and their possible interaction with nuclear genomes.

We can hypothesize that viral selfish elements may have colonized the mitochondrial genome in male bivalves promoting its segregation into primordial germ cells, thus allowing the transmission to next generations and leading to DUI achievement. If this is true, the insertion event and the appearance of DUI might be causally linked, and some implications on the origin and evolution of DUI become evident. DUI presents a scattered distribution in bivalves, and two main hypotheses have been proposed so far to account for this: 1) an unique ancient origin and subsequent reversion to standard maternal inheritance in some lineages, or 2) multiple independent origins during bivalve evolution. If these novel ORFs are in some way linked to DUI establishment, a multiple origin of DUI should not be discarded, even if it is in contrast to the mostly accepted evolutionary scenario of a single origin of DUI (Zouros 2012). The overall function similarity among all analyzed ORFs supports their origin from elements of the same kind, but the impossibility to obtain a comprehensive good alignment and their conservation only among close relative species may indicate that either they originated from independent events or their fast evolution wiped out sequence similarities. Both hypotheses cannot be definitely accepted or discarded.

Finally, the general mechanism proposed above for the transmission of selfish elements would imply that bivalves are in some way prone to viral integration in the mitochondrial genome and therefore in DUI establishment, and maybe that other animals can have experienced such kind of mitochondrial transmission modification but no evidence has been found so far.

Supplementary Material

Supplementary materials S1 and S2, tables S1–S19, and figures S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Eleonora Sparnanzoni for her precious help in lab work. This work was supported by the Italian Ministero dell'Università e della Ricerca Scientifica funding (PRIN09) and by the Donazione Canziani bequest.

Literature Cited

Abad AR, Mehrtens BJ, Mackenzie SA. 1995. Specific expression in reproductive tissues and fate of a mitochondrial sterility-associated protein in cytoplasmic male-sterile bean. Plant Cell 7:271–285.

Andersson SG, et al. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140.

Aprea S, et al. 2006. Tubulin-mediated binding of human immunodeficiency virus-1 Tat to the cytoskeleton causes proteasomal-dependent degradation of microtubule-associated protein 2 and neuronal damage. J Neurosci. 26:4054–4062.

Belshaw R, et al. 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A. 101:4894–4899.

Bernsel A, Viklund H, Hennerdal A, Elofsson A. 2009. TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res. 37: W465–W468.

Bilewitch JP, Degnan SM. 2011. A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. BMC Evol Biol. 11:228.

Birky CW Jr. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet. 35: 125–148.

Boeke JD, Stoye JP. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus H, editors. Retroviruses. Plainview (NY): Cold Spring Harbor Laboratory Press. p. 343–435.

Boore JL. 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27: 1767–1780.

Bouaouina M, et al. 2012. A conserved lipid-binding loop in the kindlin FERM F1 domain is required for kindlin-mediated αIIbβ3 integrin coactivation. J Biol Chem. 287:6979–6990.

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.

Breton S, et al. 2011a. Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? Mol Biol Evol. 28:1645–1659.

Breton S, et al. 2011b. Evidence for a fourteenth mtDNA-encoded protein in the female-transmitted mtDNA of marine mussels (Bivalvia: Mytilidae). PLoS One 6:e19365.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. Science 165:349–357.

Bryant C, Fitzgerald KA. 2009. Molecular mechanisms involved in inflammasome activation. Trends Cell Biol. 19:455–464.

Butow RA, Narayan GA. 2004. Mitochondrial signaling: the retrograde response. Mol Cell. 14:1–15.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2: 393–409.

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

Charlebois RL, Clarke GD, Beiko RG, St Jean A. 2003. Characterization of species-specific genes using a flexible, web-based querying system. FEMS Microbiol Lett. 225:213–220.

Chen D, Wang M, Zhou S, Zhou Q. 2002. HIV-1 Tat targets microtubules to induce apoptosis, a process promoted by the pro-apoptotic Bcl-2 relative Bim. EMBO J. 21:6801–6810.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826.

Claverie JM, et al. 2009. Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. J Invertebr Pathol. 101:172–180.

Cogswell AT, Kenchington EL, Zouros E. 2006. Segregation of sperm mitochondria in two- and four-cell embryos of the blue mussel *Mytilus edulis*: implications for the mechanism of doubly uniparental inheritance of mitochondrial DNA. Genome 49:799–807.

Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448: 105–114.

Coscoy L, Ganem D. 2003. PHD domains and E3 ubiquitin ligases: viruses make the connection. Trends Cell Biol. 13:7–12.

Cuconati A, White E. 2002. Viral homologs of BCL-2: role of apoptosis in the regulation of virus infection. Genes Dev. 16:2465–2478.

Daubin V, Ochman H. 2004a. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res. 14:1036–1042.

Daubin V, Ochman H. 2004b. Start-up entities in the origin of new genes. Curr Opin Genet Dev. 14:616–619.

Di Tommaso P, et al. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39: W13–W17.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 13:2213–2219.

Doorbar J, Ely S, Sterling J, McLean C, Crawford L. 1991. Specific interaction between HPV-16 E1-E4 and cytokeratins results in collapse of the epithelial cell intermediate filament network. Nature 352:824–827.

Douglas AE, Corbett KD, Berger JM, McFadden G, Handel TM. 2007. Structure of M11L: a myxoma virus structural homolog of the apoptosis inhibitor, Bcl-2. Protein Sci. 16:695–703.

Egelé C, et al. 2008. Modulation of microtubule assembly by the HIV-1 Tat protein is strongly dependent on zinc binding to Tat. Retrovirology 5:62.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 9:397–405.

Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 13:283–296.

Fischer D, Eisenberg D. 1999. Finding families for genomic ORFans. Bioinformatics 15:759–762.

Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res. 117:5–16.

Forterre P. 2010. The universal tree of life and the Last Universal Cellular Ancestor: revolution and counterrevolutions. In: Caetano-Annolés G, editor. Evolutionary genomics and systems biology. Hoboken (NJ): John Wiley and Sons, Inc. p. 58–62.

Fujii S, Toriyama K. 2008. Genome barriers between nuclei and mitochondria exemplified by cytoplasmic male sterility. Plant Cell Physiol. 49: 1484–1494.

Galluzzi L, Brenner C, Morselli E, Touat Z, Kroemer G. 2008. Viral control of mitochondrial apoptosis. PLoS Pathog. 4:e1000018.

Gestal C, et al. 2008. Study of diseases and the immune system of bivalves using molecular biology and genomics. Rev Fish Sci. 16:133–156.

Ghiselli F, et al. 2013. Structure, transcription and variability of metazoan mitochondrial genome. Perspectives from an unusual mitochondrial inheritance system. Genome Biol Evol., Advance Access published July 23, 2013, doi:10.1093/gbe/evt112.

Gissi C, Iannelli F, Pesole G. 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity 101:301–320.

Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. 2000. The origins and ongoing evolution of viruses. Trends Microbiol. 8:504–508.
Hiller K, Grote A, Scheer M, Münch R, Jahn D. 2004. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res. 32: W375–W379.
Hofmann K, Stoffel W. 1993. TMbase—a database of membrane spanning proteins segments. Biol Chem Hoppe-Seyler. 374:166.
Howard MB, Ekborg NA, Taylor LE, Hutcheson SW, Weiner RM. 2004. Identification and analysis of polyserine linker domains in prokaryotic proteins with emphasis on the marine bacterium *Microbulbifer degradans*. Protein Sci. 13:1422–1425.
Ireton K, Cossart P. 1997. Host-pathogen interactions during entry and actin-based movement of *Listeria monocytogenes*. Annu Rev Genet. 31:113–138.
Juhala RJ, et al. 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol. 299:27–51.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.
Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338: 1027–1036.
Kayal E, et al. 2011. Evolution of linear mitochondrial genomes in medusozoan cnidarians. Genome Biol Evol. 4:1–12.
Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. Biol Direct. 1:29.
Kuwana T, et al. 2002. Bid, Bax, and lipids cooperate to form supramolecular openings in the outer mitochondrial membrane. Cell 111: 331–342.
Kvansakul M, et al. 2007. A structural viral mimic of prosurvival Bcl-2: a pivotal role for sequestering proapoptotic Bax and Bak. Mol Cell 25: 933–942.
Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.
McFadden CS, Sánchez JA, France SC. 2010. Molecular phylogenetic insights into the evolution of Octocorallia: a review. Integr Comp Biol. 50:389–410.
McFadden CS, van Ofwegen LP. 2013. A second, cryptic species of the soft coral genus *Incrustatus* (Anthozoa: Octocorallia: Clavulariidae) from Tierra del Fuego, Argentina, revealed by DNA barcoding. Helgol Mar Res. 67:137–147.
Mi S, et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789.
Milani L, Ghiselli F, Maurizii MG, Passamonti M. 2011. Doubly uniparental inheritance of mitochondria as a model system for studying germ line formation. PLoS One 6:e28194.
Milani L, Ghiselli F, Passamonti M. 2012. Sex-linked mitochondrial behavior during early embryo development in *Ruditapes philippinarum* (Bivalvia Veneridae) a species with the Doubly Uniparental Inheritance (DUI) of mitochondria. J Exp Zool B Mol Dev Evol. 318: 182–189.
Morse DE, Duncan H, Hooker N, Morse A. 1977. Hydrogen peroxide induces spawning in molluscs, with activation of prostaglandin endoperoxide synthetase. Science 196:298–300.
Mouhamadou B, Barroso G, Labarère J. 2004. Molecular evolution of a mitochondrial *polB* gene, encoding a family B DNA polymerase, towards the elimination from *Agrocybe* mitochondrial genomes. Mol Genet Genomics. 272:257–263.
Nakagawa S, et al. 2013. Dynamic evolution of endogenous retrovirus-derived genes expressed in bovine conceptuses during the period of placentation. Genome Biol Evol. 5:296–306.
Nizampatnam NR, Harinath D, Yamini KN, Sujatha M, Dinesh Kumar V. 2009. Expression of sunflower cytoplasmic male sterility-associated open reading frame, *orf*H522 induces male sterility in transgenic tobacco plants. Planta 229:987–1001.
Nutt LK, et al. 2002. Bax and Bak promote apoptosis by modulating endoplasmic reticular and mitochondrial Ca2+ stores. J Biol Chem. 277:9219–9225.
Ogata H, et al. 2011. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. Intl Soc Microbial Ecol J. 5:1143–1151.
Ohta A, Nishiyama Y. 2011. Mitochondria and viruses. Mitochondrion 11: 1–12.
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.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.
Pons J-L, Labesse G. 2009. @TOME-2: a new pipeline for comparative modeling of protein–ligand complexes. Nucleic Acids Res. 37: W485–W491.
Pont-Kingdon G, et al. 1995. A coral mitochondrial MutS gene. Nature 375:109–111.
Rohwer F. 2003. Global phage diversity. Cell 113:141.
Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. Bioinformatics methods and protocols: methods in molecular biology. Totowa (NJ): Humana Press. p. 365–386.
Shao Z, Graf S, Chaga OY, Lavrov DV. 2006. Mitochondrial genome of the moon jelly *Aurelia aurita* (Cnidaria, Scyphozoa): a linear DNA molecule encoding a putative DNA-dependent DNA polymerase. Gene 381: 92–101.
Siew N, Fischer D. 2004. Structural biology sheds light on the puzzle of genomic ORFans. J Mol Biol. 342:369–373.
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, et al. 2011. First complete mitochondrial genome sequence from a box jellyfish reveals a highly fragmented linear architecture and insights into telomere evolution. Genome Biol Evol. 4: 52–58.
Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33:W244–W248.
Soltys BJ, Gupta RS. 2000. Mitochondrial proteins at unexpected cellular locations: export of proteins from mitochondria from an evolutionary perspective. Int Rev Cytol. 194:133–196.
Suttle CA. 2005. Viruses in the sea. Nature 437:356–361.
Takeuchi O, Akira S. 2009. Innate immunity to virus infection. Immunol Rev. 227:75–86.
Tamura K, et al. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.
Todd AE, Orengo CA, Thornton JM. 2001. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol. 307: 1113–1143.
Wang HW, Sharp TV, Koumi A, Koentges G, Boshoff C. 2002. Characterization of an anti-apoptotic glycoprotein encoded by Kaposi's sarcoma-associated herpesvirus which resembles a spliced variant of human survivin. EMBO J. 21:2602–2615.
Wei MC, et al. 2001. Proapoptotic BAX and BAK: a requisite gateway to mitochondrial dysfunction and death. Science 292:727–730.

White E, Cipriani R, Sabbatini P, Denton A. 1991. Adenovirus E1B 19-kilodalton protein overcomes the cytotoxicity of E1A proteins. J Virol. 65:2968–2978.
Woodson JD, Chory J. 2008. Coordination of gene expression between organellar and nuclear genomes. Nat Rev Genet. 9:383–395.
Wu X, et al. 2012. New features of Asian *Crassostrea* oyster mitochondrial genomes: a novel alloacceptor tRNA gene recruitment and two novel ORFs. Gene 507:112–118.
Xu J, Zhang Y. 2010. How significant is a protein structure similarity with TM-score=05?. Bioinformatics 26:889–895.
Yin Y, Fischer D. 2006. On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC Evol Biol. 6:63.
Yu G, Stoltzfus A. 2012. Population diversity of ORFan genes in *Escherichia coli*. Genome Biol Evol. 4:1176–1187.
Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.
Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.
Zouros E. 2012. Biparental inheritance through uniparental transmission: the doubly uniparental inheritance (DUI) of mitochondrial DNA. Evol Biol. 40:1–31.
Zouros E, Oberhauser Ball A, 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*. Proc Natl Acad Sci U S A. 91:7463–7467.

Associate editor: Bill Martin