Université de Montréal

Caractérisation bioinformatique des nouvelles protéines mitochondriales chez les moules d'eau douce (Bivalvia : Unionoida).

par
Alyssa Mitchell

Département de sciences biologiques
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de M.Sc.
en sciences biologiques

Décembre, 2015

Résumé

Malgré que le contenu des génomes mitochondriaux animaux soit dit bien conservé, des nouveaux gènes mitochondriaux ont été identifiés chez plusieurs espèces, surtout des invertébrés. Par exemple, les bivalves exhibant la double transmission uniparentale de leurs génomes mitochondriaux possèdent des nouveaux gènes spécifiques au sexe (M-*ORF* dans l'ADN de type M, F-*ORF* dans l'ADN de type F) qui ont été caractérisés *in silico* chez trois espèces de l'ordre Mytiloida, une espèce de Veneroida et une espèce de Unionoida par une précédente étude. Même si les séquences varient beaucoup entre ces trois ordres, cette étude à montré que des hélices transmembranaires ainsi que des peptides signaux sont conservés pour toutes les séquences. L'étude a aussi montré que les nouveaux gènes pourraient avoir des rôles dans la signalisation cellulaire, le cycle cellulaire et la réponse immunitaire et qu'ils pourraient être le résultat de l'endogénisation de l'ADN viral. Le projet présenté ici a pour but de mieux caractériser ces nouveaux gènes et leur origine potentielle, en plus d'étudier le H-*ORF* particulier aux hermaphrodites, en ciblant les espèces des unionidés. Les résultats montrent que les hélices transmembranaires et peptides signaux sont conservés chez les unionidés, les protéines semblent être associées à la membrane et être capables de lier des acides nucléiques et protéines, et les fonctions potentielles sont conservées. Les M-*ORFs* semblent avoir un rôle dans le transport et des processus cellulaires tels que la signalisation, le cycle cellulaire et la division, et l'organisation du cytosquelette. Les F-*ORFs* semblent être impliqués dans le trafic et transport cellulaire et la réponse immunitaire. Finalement, les H-*ORFs* semblent être des glycoprotéines structurales avec des rôles dans la signalisation, le transport et la transcription. Les résultats de ce projet pourraient supporter une origine virale ou mitochondriale pour ces gènes.

Mots-clés : double transmission uniparentale, Unionoida, ORFans, mitochondrie, *in silico*

Abstract

Although animal mitochondrial gene content is generally considered to be well-conserved, new genes have been identified in a variety of species, particularly invertebrates. For example, bivalves with doubly uniparental inheritance (DUI) of the mitochondrial genome have novel, sex-specific genes (M-*ORF* in M-type DNA, F-*ORF* in F-type DNA) which have been characterized *in silico* in three species of the order Mytiloida, one Veneroida and one Unionoida in a previous study. Although they are highly variable across these three orders, this study found conserved N-terminal signal peptides and transmembrane helices across all species. The study also showed that the new genes may and have roles in cell signaling, cell cycle, and immune response, and that they may be the result of endogenization of viral DNA. This project aimed to better characterize these novel genes and their potential origin as well as the H-*ORF* specific to hermaphrodites by focusing on the Unionoida. The pattern of conserved transmembrane helices and signal peptides is present across the species studied, all proteins seem to be membrane associated and able to bind nucleic acids and proteins, and potential functions are conserved as well. M-*ORFs* seem to have a role in transport and cellular processes such as signalling, cell cycle and division, and cytoskeleton organisation. F-*ORFs* are predicted to be involved in cellular traffic and transport and immune response. Finally, H-*ORFs* appear to be structural glycoproteins which may be involved in signalling, transport and transcription. The results of this project support either a viral or a mitochondrial origin for these genes.

Keywords : doubly uniparental inheritance, Unionoida, ORFans, mitochondria, *in silico*

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations et sigles

ABC : transporteur ABC / ATP-binding cassette transporter

ADN : acide désoxyribonucléique / deoxyribonucleic acid

ADNmt : ADN mitochondriale / mitochondrial DNA

ARN : acide ribonucléique / ribonucleic acid

ATP : adénosine triphosphate / adenosine triphosphate

CMS : stérilité cytoplasmique mâle / cytoplasmic male sterility

CTERM : C-terminale / C-terminal

DUI : double transmission uniparentale / doubly uniparental inheritance

e.g. : exempli gratia

Et al. : et alii

HMM : modèle de Markov caché / Hidden Markov Model

i.e. : id est

MY : million d'années / million years

NADH : nicotinamide adénine dinucléotide / nicotinamide adenine dinucleotide

ORF : cadre de lecture ouvert / open reading frame

ORFan : cadre de lecture ouvert sans homologie à une protéine connue / open reading frame without homology to a known protein

PPR : protéines avec répétitions de type pentatricopeptide / pentatricopeptide repeat proteins

SMI : transmission strictement maternelle / strict maternal inheritance

SP : peptide signal / signal peptide

Spp. : espèces / species

TMH : hélice transmembranaire / transmembrane helix

UPR_{mt} : réponse au stress lié à l'accumulation de protéines mal repliées dans les mitochondries / mitochondrial unfolded protein response

For all those who part ways by quietly chanting the name of a professor who looks like Santa.

Hockey hamster.

Remerciements

- Je commence par Sophie, bien sûr! Surtout pour la décision aventureuse de m'accepter comme première étudiante graduée dans ton labo pas-tout-à-fait-existant, mais aussi pour les cent mille lettres de référence, les soupers, les daiquiris sans alcool et les cent mille rires.

- Don, I think you're the hardest to fit into a bullet point. You took me on as an honours student at the absolute last minute when no one knew much of anything about me (or even who I was). The four years you've known me have been a bit chaotic to say the least, but you've been patient and supportive through it all (I deleted "very" for you). Coincidentally, your name means "gift" in French. If I ever master my great-grandmother's molasses cookies, you're getting a great big batch!

- Of course I have to thank my family, aka the reason I exist. I think I've planned the perfect celebratory feast: cheeseball, lobster dip, almond roca, gingerbread men from Yarmouth, and braid. Nobody needs a main course. Make it a kitchen party, and get ready to atone for your sins at the lobster dip (*forgiiive me Father, I know not what I do!*).

- Next is my nerd crew (plus that one "normal" guy with the tin foil hat and hatchet). Every now and then I get a bit stunned that we're all still in touch and see each other anywhere near as often as we do. I think that's a pretty good reflection of how highly these friendships are prioritized, and it's nice to know that we're pretty ride or die (or maybe hike or die). I'm thrilled to have found my people so early in life, and when I'm home for Christmas I want to celebrate this like it's The Night Pat Murphy Died.

- Je tiens aussi à remercier mon département, surtout ceux et celles que j'ai connu le plus. D'abord, ce « grand lab avec beaucoup d'étudiants » à Bernard Angers, qui m'ont adoptée ma première année. Trouvez-vous que mes présentations ont évolué au cours de ma maitrise? Ensuite, mon labo, qui a

commencé à se peupler tout tranquillement – c'est plus fun en groupe qu'à 1 ou 2. Il faut que je rémercie les étudiants au bac avec qui j'ai travaillé et parmi lesquels je compte plusieurs amis - vous avez mis beaucoup de joie dans les TPs et séances de tutorat! Finalement, les démos et tuteurs qui ont travaillé à côté de moi, et tout le gang qui se réunit au CI. C'est grâce à vous que j'ai eu l'expérience d'immersion totale que j'ai voulue, et c'est aussi grâce à vous que j'ai pu rire au moins une fois par jour, même quand ça n'allait pas bien.

- *Gli italiani* are next, aka the Passamonti lab. *Grazie di* your contributions to my work, the article that guided mine, your contributions to our lab (Davide and Stefano), and the entertaining and delicious visits from Dr. Over-the-hill. *Molte grazie!*
- For beach days, Nordic spas, masala dosa, all of the most useful Hindi vocabulary (food), and sternly telling me that I need to meditate, शुक्रिया राजेश.
- I'm going to close this section by thanking all of the artists who made the most miserable tasks a bit more fun. For all the hours of spent sorting hits, scoring my beastly matrix, and moving pieces of figures one pixel at a time, thank you Prince Royce, Taylor Swift, Lea Michele, Tim Chaisson, and Great Big Sea. For the writing process, thank you Mozart, Beethoven, and Bach.

Chapitre 1 : Introduction générale

Section 1.1 : La mitochondrie et son génome

Les mitochondries sont des organelles à double-membrane retrouvées dans le cytoplasme chez les cellules eucaryotes et responsables de la production d'énergie. Bien que ça soit leur fonction principale, elles sont également impliquées dans d'autres processus cellulaires, tels que la signalisation cellulaire, la régulation métabolique, le contrôle du cycle cellulaire, le développement, la réponse antivirale et l'apoptose [1].

Antérieurement, on croyait que les génomes mitochondriaux (ADN mitochondrial ou ADNmt) chez les espèces animales étaient tous très similaires – une molécule circulaire d'environ 15 000 à 20 000 paires de bases encodant 2 ARN ribosomaux, 22 ARN de transfert, et 13 protéines impliquées dans la synthèse de l'ATP à l'intérieur des mitochondries [7 sous-unités du complexe de la NADH-ubiquinone oxydoréductase (gènes *nad1-6*, *nad4L*), une du complexe de l'ubiquinol-cytochrome c oxydoréductase (gène *cytb*), 3 du complexe de la cytochrome *c* oxydase (gènes *cox1-3*), et 2 du complexe de l'ATP synthase (gènes *atp6* et *atp8*)] [2, 3]. Aujourd'hui on connaît plusieurs cas de réarrangements structuraux du génome mitochondrial ou encore la présence de gènes mitochondriaux supplémentaires (Voir [4] pour une revue). Par exemple, les Medusozoa ont un génome mitochondrial linéaire [5], et certains crustacés terrestres ont une portion circulaire et une portion linéaire [6]. Au niveau des gènes, on connait plusieurs cas de duplications de gènes existants ou encore de découverte de nouveaux gènes, surtout chez les invertébrés [4]. Chez les bivalves en particulier, deux exemples de duplication de gènes codant pour des protéines sont bien connus : la duplication de *cox2* chez *Musculista senhousia* [7], et la duplication de *nad2* chez le genre *Crassostrea* [8]. Le gène codant l'ARN ribosomal rrnS est aussi dupliqué chez *Crassostrea gigas* [3, 9]. Chez *Aurelia aurita,* et le genre *Pocillopora* (Cnidaria), un nouveau cadre de lecture (« *open reading frame* » ou ORF) de fonction inconnue a été retrouvé dans le génome mitochondrial [10–12] et deux ORFs de fonctions inconnues existent également chez *Iphitheon panicea* (Porifera) [13].

De plus, la transmission strictement maternelle (« *Strict maternal inheritance* » ou SMI), qui est la norme dans le règne animal, était considérée le seul système de transmission mitochondriale [14]. On connait maintenant une exception à cette règle aussi – la double transmission uniparentale, qui sera décrite dans la prochaine section.

Section 1.2 : La double transmission uniparentale

Il existe un cas exceptionnel à la « règle » de la transmission strictement maternelle de l'ADNmt chez les animaux. Plusieurs espèces de bivalves (Ordres Mytiloida, Nuculanoida, Unionoida et Veneroida) ont un mode de transmission fondamentalement différent connu sous le nom de « double transmission uniparentale » ou DUI (« *Doubly uniparental inheritance »*) [15–22]. Ces espèces sont caractérisées par la présence de deux ADNmt distincts : un génome M transmis par les mâles, et un génome F transmis par les femelles. Normalement, un œuf haploïde contient seulement des mitochondries de type F (voir [19, 20] pour des exceptions), et les spermatozoïdes contiennent seulement des mitochondries de type M, qui vont entrer dans l'œuf lors de la fécondation. Chez les embryons destinés à devenir femelles, les mitochondries paternelles sont dispersées dans toutes les cellules de l'embryon, et sont détruites pour permettre le développement d'une femelle homoplasmique (l'homoplasmie – où toutes les copies d'ADNmt sont identiques – est la norme sous SMI). Chez les embryons destinés à devenir mâles, par contre, les mitochondries paternelles sont regroupées dans les cellules destinées à devenir la gonade (Figure 1). Un mâle mature est hétéroplasmique, avec l'ADNmt de type F dans ses tissus somatiques, et l'ADNmt de type M dans sa gonade [21–23] (cet ADNmt M est principalement actif dans les spermatozoïdes [24, 25]).

Le taux de divergence entre les deux génomes au niveau des nucléotides peut varier d'environ 10% chez les moules marines à plus de 50% chez les moules d'eau douce, et une évolution plus rapide du génome mt de type M a été notée chez la plupart des espèces [21, 29-33]. Une explication possible pour cette différence est que le génome mt de type M serait un élément égoïste (ou « presque égoïse ») puisqu'il est fonctionnel seulement dans les cellules spermatogéniques [28]. D'autres études ont également proposé que les deux génomes évolueraient de façon neutre, mais que le type M accumulerait des mutations plus rapidement

dû à (i) sa population effective plus petite, (ii) son taux de réplication et donc de mutations élévé (on observe un total de sept divisions pendant la gamètogénèse chez le mâle versus 4 chez la femelle), ou (iii) aux dommages oxidatifs plus importants que les mitochondries subissent chez les spermatozoïdes [e.g. 21].

Deux hypothèses non-exclusives ont été proposées pour expliquer l'origine et le maintien du système atypique DUI chez les bivalves [29, 30] : (i) l'ADNmt mâle est impliqué dans des fonctions spécifiques et nécessaires aux spermatozoïdes et/ou (ii) ces deux génomes associés aux sexes sont impliqués dans la détermination du sexe. La détermination du sexe est méconnue chez les bivalves, mais il y a une particularité connue chez les Mytiloida et Veneroida – un effet maternel sur la sexe-ratio [36-38]. Chaque femelle produit des descendants majoritairement femelles, majoritairement mâles, ou environ 50% femelles et 50% mâles, peu importe avec quel mâle elle a été croisée. De plus, la majorité des filles présentent le même biais que la mère, mais des changements de biais qui suivent un rapport Mendelien ont également été observés. Il a été suggéré qu'un facteur nucléaire maternel clé ainsi que des facteurs secondaires nucléaires et/ou mitochondriaux serait impliqués dans le maintient de ces sexe-ratios biaisées [31].

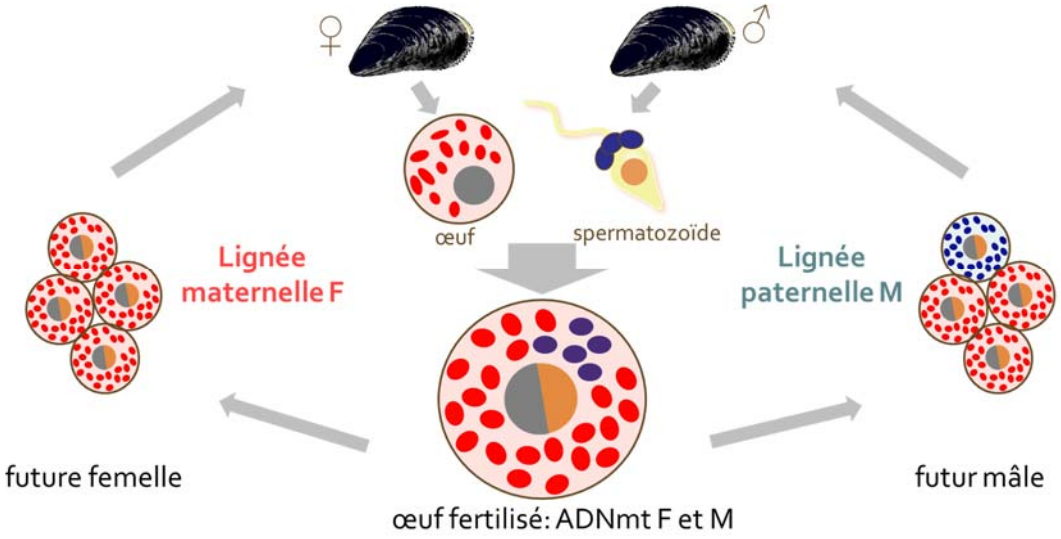


Figure 1. La double transmission uniparentale. Mitochondries avec génome de type M en bleu; mitochondries avec génome de type F en rouge.

Section 1.3 : Les moules d'eau douce (Unionoida)

Les moules d'eau douce (Unionoida) sont un groupe relativement ancien, avec une origine durant le Triassique il y a plus de 200 millions d'années [32, 33]. Les espèces sont généralement gonochoriques (à sexes séparés), avec un cycle de vie fondamentalement différent de celui des moules marines Mytiloida et palourdes marines Veneroida qui relâchent leurs gamètes dans la colonne d'eau, où a lieu la fécondation et le développement larvaire. Chez les Unionoida, seuls les mâles relachent leurs gamètes dans l'eau, les spermatozoïdes sont captés par les femelles et les premiers stades de développement ont lieu dans des compartiments spécialisés appelés marsupium dans les branchies des femelles [34]. Quand les embryons atteignent un certain stade de développement, la femelle les relâche, et ils doivent s'encyster sur les branchies d'un poisson hôte pour vivre une métamorphose parasitique. Une fois transformés en juvéniles, ils se détachent et tombent au fond du cours d'eau (pour plus de détails, voir [34])

Un petit nombre d'espèces hermaphrodites a été reporté chez les Unionoida (*e.g.*, [32, 35, 36]). En Amérique du Nord, par exemple, seulement 7 espèces sont hermaphrodites sur plus de 300 espèces repertoriées [37]. De rares individus hermaphrodites sont également périodiquement trouvés chez les espèces à sexes séparés [37]. Plusieurs hypothèses ont été proposées pour expliquer l'émergence de l'hermaphrodisme chez les Unionoida, comme par exemple des facteurs environnementaux tels que la force du courant, la position d'un individu dans la population pour donner ou capter des spermatozoïdes et la densité de la population, e.g. [38]) mais cela demeure encore nébuleux [39]. Aussi, il n'est pas connu si des différences génétiques existent entre les hermaphrodites obligatoires et accidentels, cependant il y a des différences anatomiques bien documentées – tous les hermaphrodites ont un ovotestis (une gonade contenant à la fois des cellules spermatogéniques et des cellules ovogéniques), mais la distribution des cellules spermatogéniques et ovogéniques diffère entre les deux types d'hermaphrodites. Chez les hermaphrodites obligatoires on observe des acini discrèts qui produisent un type de gamète, mais chez les hermaphrodites accidentels on observe une distribution aléatoire de ces cellules [37]. Des analyses phylogénétiques démontrent que les espèces hermaphrodites en Amérique de Nord sont relativement jeunes [32, 37], et que l'hermaphrodisme est un caractère dérivé qui évolue à partir des femelles ([40, 41], voir Figure 2).

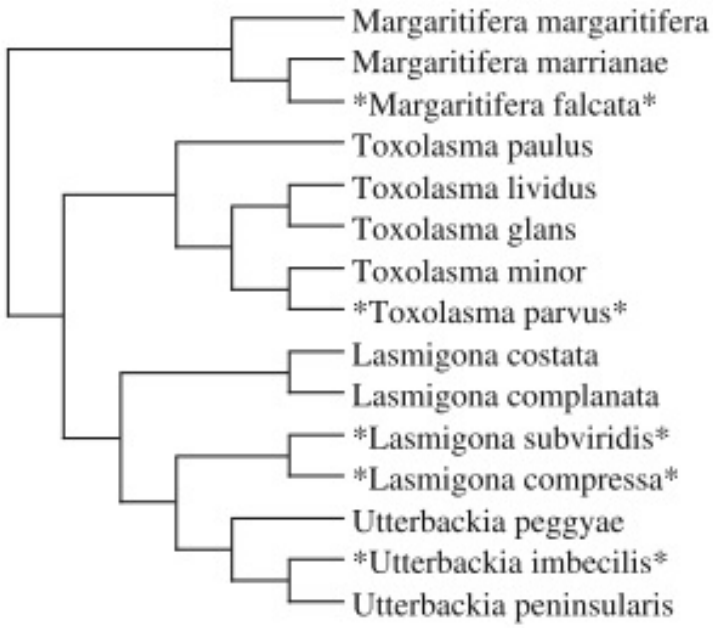


Figure 2. Phylogénie simplifiée d'une collection d'espèces à sexes séparés et hermaphrodites (familles Unionoida et Margaritifera). Les espèces marquées par une étoile (*) sont des hermaphrodites qui ont perdu le génome mitochondrial de type M (donc possédant une transmission mitochondriale strictement maternelle). Figure tirée de Stewart et al. [41] avec permission.

Section 1.4 : Nouveautés et nouveaux gènes

Récemment, Breton et al. [40, 42] ont identifié des nouveaux gènes codant pour des protéines dans les génomes mitochondriaux des moules d'eau douce. Tel que mentionné précédemment, chez les unionidés, la grande majorité des espèces ont des sexes séparés mais il existe aussi des rares cas d'espèces hermaphrodites [32, 39]. Breton et al. [40] ont séquencé les génomes mitochondriaux des mâles et femelles pour plusieurs espèces gonochoriques, ainsi que le génome mitochondrial présent chez 5 espèces hermaphrodites proches parentes des espèces gonochoriques, mais qui ont toutes évolué de façon indépendante. Dans chacun des génomes, un quatorzième ORF a donc été découvert (F-*ORF* dans les génomes F, M-*ORF* dans les génomes M et H-*ORF* dans les génomes des espèces hermaphrodites) [40, 42]. La

technique Western Blot a été utilisée pour démontrer que les F-*ORF* et M-*ORF* sont exprimés, par contre, cela reste encore à être démontré pour le H-*ORF*. Les études de Breton et al. [40, 42] ont présenté six points importants :

i. tous les génomes F étudiés contiennent un F-*ORF* codant pour une protéine qui est conservée entre les espèces, mais qui ne présente aucune homologie évidente aux autres protéines connues (selon les résultats BLAST);

ii. tous les génomes M étudiés contiennent un M-*ORF* codant pour une protéine qui est conservée entre les espèces, mais qui n'est pas homologue au F-*ORF*, ni à aucune protéine connue;

iii. les hermaphrodites n'ont pas de génome M et leur ADNmt contient un F-*ORF* hautement modifié (H-*ORF*);

iv. les H-*ORF*s divergent des séquences F-*ORF* de leurs espèces proches parentes - leurs séquences sont plus longues (environ 80 acides aminés pour les F-*ORF*s et 150 acides aminés pour les H-*ORF*s) et elles possèdent plusieurs sous-unités répétitrices et plusieurs portions transmembranaires prédites versus une seule chez les F-*ORF*s et M-*ORF*s;

v. le gonochorisme est toujours accompagné de la DUI et la présence des F-*ORF*s et M-*ORF*s tandis que l'hermaphrodisme est accompagné de la SMI et la présence d'un F-*ORF* hautement modifié (H-*ORF*), ce qui mène à l'hypothèse que la DUI et les nouveaux gènes auraient un rôle dans le maintien des sexes séparés chez les unionidés;

vi. l'analyse immunohistochimique indique que la protéine encodée par le F-*ORF* chez l'espèce *Venustaconcha ellipsiformis* est non seulement présente dans la mitochondrie, mais transportée hors de l'organelle et retrouvée dans la membrane nucléaire et le nucléoplasme des œufs;

Cette dernière observation indique un rôle autre que la phosphorylation oxydative, et des études subséquentes sont venues appuyer l'hypothèse que ce produit de gène pourrait jouer un rôle dans la détermination du sexe [40, 42–45]. Présentement, il n'y a aucun cas connu dans le règne animal où les mitochondries sont directement impliquées dans la détermination

du sexe. Les fonctions des nouveaux gènes mitochondriaux découverts chez les bivalves avec la DUI demeurent pour le moment obscures (voir ci-dessous). Les modifications importantes observées dans le gène H-*ORF* chez les espèces hermaphrodites suggèrent une fonction différente pour ce gène ou encore une perte de fonction, mais cela reste à être étudié. Le lien entre la DUI et la détermination du sexe reste aussi à être élucidé, et la raison pour la déviation de la SMI chez les bivalves gonochoriques – et un retour vers la SMI chez les hermaphrodites – demeurent des questions ouvertes.

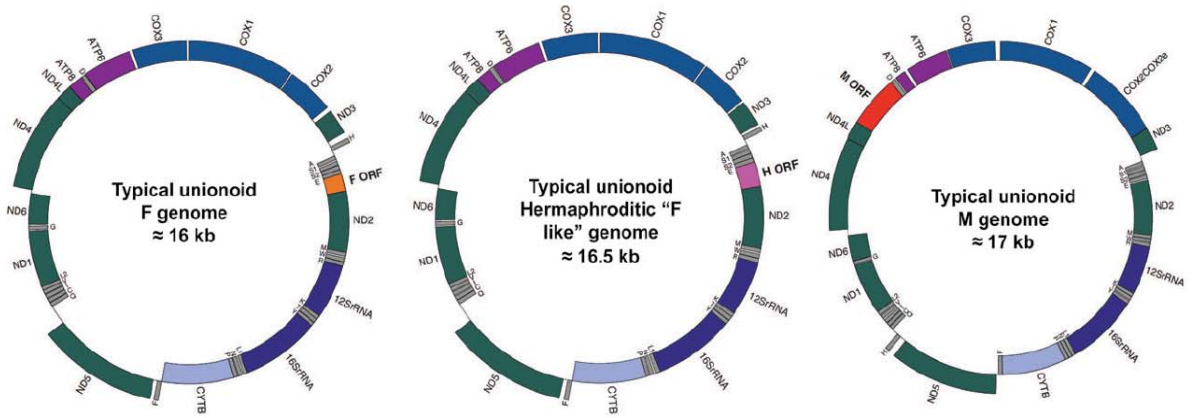


Figure 3. Cartes des génomes de types F, H et M. Identités des gènes : complexe I en vert; complexe III en bleu pâle, complexe IV en bleu; complexe V en violet; ARN ribosomaux en bleu foncé. Les ARNs de transfert sont indiqués par leur lettre d'acide aminée. F-*ORF*, orange; H-*ORF*, rose; M-*ORF*, rouge. Les gènes à l'intérieure du cercle sont encodés sur le brin léger, ceux à l'extérieure du cercle sont encodés sur le brin lourd. Figure et légende tirées de Breton et al. 2011 [40] avec permission.

Milani et al. [44] ont publié les premières analyses *in silico* des structures et fonctions potentielles des F-*ORFs* et M-*ORFs* chez les bivalves avec la DUI (les moules marines *Musculista senhousia* and *Mytilus* spp. (Mytiloida), la palourde marine *Ruditapes philippinarum* (Veneroida), et l'unionidé *Venustaconcha ellipsiformis* (Unionoida)). Leurs résultats ont montré que les séquences des deux nouveaux ORFs sont très variables au niveau

des nucléotides et des acides aminés, et que ces gènes évoluent plus vite que tout autre gène mitochondrial chez les bivalves étudiés [40].

Les prédictions structurales et fonctionnelles indiquaient des similarités parmi toutes les espèces. D'abord, des hélices transmembranaires et peptides signaux sont conservés entre les espèces. Ensuite, les fonctions prédites pour les F-*ORFs* incluent la liaison avec les acides nucléiques, l'association aux membranes pour la signalisation ou l'adhésion cellulaire, ou un rôle dans la réponse immunitaire, tandis que les fonctions prédites pour les M-*ORFs* incluent l'association aux membranes, des interactions avec les acides nucléiques (surtout pour la signalisation cellulaire et la différentiation et développement), des interactions avec le cytosquelette, l'ubiquitination, l'apoptose et la réponse immunitaire [44].

En plus de prédire la structure et la fonction de ces protéines, cette étude a émis une hypothèse sur leur origine : l'endogénisation d'un ADN viral [44, 46]. Toutefois, en raison des taxons étudiés (5 mytilidés, un veneridé, et un unionidé) qui sont évolutivement très distants, et en raison des problèmes pour l'obtention de bons alignements des séquences, les auteurs ont également émis l'hypothèse que les ORFs chez les bivalves avec la DUI pourraient provenir d'événements d'endogénisation indépendants [17, 18]. Une augmentation du nombre d'espèces proches parentes et de séquences à l'étude pourrait aider à avoir une meilleure idée de l'origine de ces nouveaux gènes mitochondriaux. Par exemple, au moins quatre autres origines peuvent être proposées : (i) un gène homologue à un ancien gène bactérien, (ii) une duplication et néofonctionalization d'un gène mitochondrial, (iii) une origine à partir de séquences mitochondriales non-codantes, et (iv) un transfert du noyau vers la mitochondrie.

Les unionidés représentent un excellent modèle pour mieux comprendre l'origine et les fonctions de ces nouveaux gènes. Ils ont une position basale dans les Bivalvia, et les séquences complètes des M-*ORF*, F-*ORF* et H-*ORF* sont disponibles pour plusieurs espèces gonochoriques et des espèces hermaphrodites proche-parentes. La gamme de séquences disponibles représentent au moins 200 MY d'évolution pour ces gènes ([40]; Guerra et al., en prep), mais les séquences demeurent relativement similaires et donc plus facilement comparables.

Section 1.5 : Objectifs et hypothèses

L'objectif principal du projet est de caractériser l'évolution, la structure et la fonction des nouveaux gènes mitochondriaux F-*ORF*, H-*ORF* et M-*ORF* chez les moules d'eau douce, à partir de leur séquences nucléotidiques et/ou protéiques. Notre hypothèse générale est que ces gènes sont impliqués dans la détermination du sexe et donc subissent de fortes pressions sélectives. Plus précisément, on prévoit que (i) ces gènes évoluent rapidement et présentent plus de mutations non-synonymes que synonymes chez toutes les espèces (une signature répandue chez les gènes impliqués dans la détermination du sexe), (ii) malgré cette évolution rapide, les structures secondaires devraient être conservées pour chacun des nouveaux gènes : c'est-à-dire entre les espèces pour les F-*ORFs* et M-*ORFs*, et intra-espèce pour les H-*ORFs*, et (iii) la fonction prédite pour chacun des nouveaux gènes devrait être la même pour toutes les espèces pour un même genre (ex. les séquences F-*ORF* donnant des résultats similaires chez toutes les espèces).

Chapitre 2 : *In silico* analyses of mitochondrial ORFans in freshwater mussels (Bivalvia: Unionoida) provide framework for future studies of their origin and function.

Article in preparation for *BMC Genomics.*

Alyssa Mitchell[a], Davide Guerra[a], Donald Stewart[b], Sophie Breton [a]

[a]Department of Biological Sciences, Université de Montréal, Montréal, QC H3C 3J7 Canada

[b]Department of Biology, Acadia University, Wolfville, NS B4P 2R6 Canada

Section 2.1: Introduction

Metazoan mitochondrial genomes (mtDNAs) are typically small, circular genomes without introns that encode 2 ribosomal RNAs, 22 transfer RNAs, and 13 proteins involved in ATP production [2, 3]. Strict maternal inheritance (SMI) of mtDNA is predominant among animals with limited or no paternal contribution [14]. There are, however, many exceptions to these characteristics. For example, linearized mitochondrial genomes have been reported in the Medusozoa [5] and some terrestrial isopod crustaceans [6]. Differences in gene content have also been found among metazoan mtDNAs, particularly in invertebrates (see [4] for a review). For example, duplications of typical protein-coding genes have been reported in several mollusc species, including cephalopods, aplacophorans, and bivalves; additional 'atypical' protein-coding genes with non-OXPHOS functions have been reported in cnidarians, sponges, and placozoans (e.g., *atp9*, *dnaB*, *tatC*); and mitochondrial ORFans, i.e., 'atypical' genes with unknown function, have been identified in cnidarians and in bivalves with doubly uniparental inheritance of mtDNA (DUI), which is the only known exception to SMI in the animal kingdom [4].

DUI has been reported in marine and freshwater bivalves (Orders Mytiloida, Nuculanoida, Unionoida, and Veneroida) ([26]; [47]; [27]; [21]). Species with DUI possess mitochondrial genomes that are transmitted in a sex-specific manner (known as female F-type

and male M-type mtDNAs, respectively). Haploid eggs typically contain mitochondria with only F-type mtDNA (but see [19] and [20]), while sperm mitochondria, which enter the egg when fertilization occurs, only contain the M-type. If the embryo develops as a female, sperm mitochondria are dispersed and/or destroyed, leading to homoplasmic females (similar to what happens under SMI). If the embryo develops as a male, sperm mitochondria remain grouped together, and are eventually sequestered in the germ line, which becomes homoplasmic for the M mtDNA [24, 25]. Males are therefore heteroplasmic individuals, with mitochondria inherited from their mother containing the F-type mtDNA throughout their soma, and mitochondria inherited from their father containing the M-type mtDNA in germ line cells (in males M mtDNA can be found in variable proportions also in somatic tissues [9]). DNA divergence between conspecific M- vs. F-type mitochondrial genomes over 50% has been found in Unionoida ([21]).

With their unique DUI system, bivalves not only challenge our traditional view of the SMI of mtDNA, their mitochondrial genomes also contain additional, sex-specific protein-coding genes, i.e., the mitochondrial ORFans - F-*orfs* and M-*orfs* in the F- and M-type mtDNAs, respectively - whose products are exported from the organelle and may be involved in functions other than energy production [40, 42–46]. In freshwater mussels, for example, species typically have separate sexes (gonochorism or dioecy), but hermaphroditism also occurs rarely [32, 39]. An absolute correlation has been observed between gonochorism and the presence of DUI and novel sex-specific proteins encoded by the F- and M-type mtDNAs (F-*ORF* and M-*ORF*), whereas hermaphroditic species lack the M-type altogether [43]. Hermaphroditic species appear to follow the SMI rule of mitochondrial transmission and individual mussels have only one type of mtDNA, called H-type [16]. The H-type is remarkably similar to the F-type mtDNA of closely-related gonochoric species except for the novel ORFan gene (named H-*orf* in these species), which is a highly mutated version of the F-*orf* in their sister taxa [43]. For these reasons, Breton et al. [43] proposed a connection between DUI and the maintenance of separate sexes in freshwater mussels. However, the link between DUI and sex determination, and the cause of deviation from the "SMI rule" in bivalves are still open questions.

The first in-depth bioinformatic analysis of the structures and potential functions of F-*ORF* and M-*ORF* proteins was performed by Milani et al. [44] on the following DUI bivalve species: the marine mussels *Musculista senhousia*, *Mytilus edulis*, *Mytilus galloprovincialis*, *Mytilus trossulus* and *Mytilus californianus* (Mytiloida), the marine clam *Ruditapes philippinarum* (Veneroida), and the freshwater mussel *Venustaconcha ellipsiformis* (Unionoida). M-*orf* and F-*orf* nucleotide sequences were found to be highly variable, with mostly non-synonymous mutations, indicating rapid evolution and supporting previous claims that these protein-coding genes are the fastest-evolving mitochondrial genes in bivalves with DUI [40, 44]. Despite this fast rate of evolution, structural similarities in their translated amino acid sequences were observed among species, and ORFan proteins were predicted to share similar functions. For example, F-*ORFs* were largely predicted to bind and interact with nucleic acids, associate with membranes for cell adhesion and/or signalling, or play a role in immune response. M-*ORFs* were also predicted to be membrane-associated and interact with nucleic acids, primarily for signalling, cell differentiation and development, and also for cytoskeleton formation and dynamics, ubiquitination, apoptosis, and immune response [44]. Even if hit probabilities were sometimes low and the regions of similarity were of short lengths, several clues suggested that these novel ORFans originated from endogenization of viral DNA [44, 46]. However, the impossibility of obtaining good alignments including F-*ORFs* and M-*ORFs* from all species, due to the highly divergent nature of the ORFans, indicated that either their fast evolution wiped out sequence similarities among species or that they originated from independent virus endogenization events [44]. It is also conceivable that the ORFans originated from different sources/processes but evolved similar function(s) in these distantly related DUI species, particularly if DUI evolved independently more than once [44]. Other than a viral origin, there are at least four other possibilities for the source of these mitochondrial ORFans; they may have originated from (i) a gene homologous to ancestral bacterial protein-coding genes, (ii) a duplicated and diverged mitochondrial gene, (iii) a gene composed from previously non-coding mitochondrial sequences, or (iv) a gene transferred from the nucleus to the mitochondrion (e.g., [40]).

The reality is that it is unfortunately not currently possible to confirm that mitochondrial ORFans in these distantly related DUI species are homologous because of their

high divergence and incomplete knowledge regarding their phylogenetic distribution. One option to better understand the origin(s) and function(s) of these ORFans is to compare more closely related sequences at a lower taxonomic level. Freshwater mussels (Unionoida) offer an excellent opportunity for this for at least two reasons: (1) the basal nature of the Unionoida within the Bivalvia, according to mtDNA-based phylogenies, suggests that their ORFans have a very ancient origin and that DUI in this group might be one of the first examples of this phenomenon in bivalves [33], and (2) complete F and M genomes or F-*orf*, M-*orf* and H-*orf* sequences are available for several gonochoric species and five independently evolved hermaphroditic species (e.g., [33, 43, 48]). All of them belong to the family Unionidae, but recently we have sequenced the F and M mtDNAs from *Cumberlandia monodonta* (Margaritiferidae) and *Hyridella menziesii* (Hyriidae) (Guerra et al., in prep), and these genomes possess an F-*orf* and an M-*orf*, suggesting that these unique genes have been present and functioning continuously for >200 million years in this group ([43]; Guerra et al., in prep).

The present study thus aims to predict the origin, structure, and function of the F-*ORF* and M-*ORF* protein sequences in Unionoida, and analyze the H-*ORFs* for the first time. Our results confirm that they are the fastest evolving genes in unionoid mitochondrial genomes, that they share structural and functional similarities, and that they may have a viral or a mitochondrial origin, bringing back on the table the evolutionary scenario of multiple origins of DUI, with the possibility of DUI systems with elements of different sources/origins and different mechanisms of action in the distantly-related DUI taxa [44, 46].

Section 2.2: Materials and methods

Section 2.2.1: Sequences used in the analyses

ORFan and *cox1* nucleotide sequences of unionoid bivalve species were either obtained from the National Center for Biotechnology Information (NCBI) or from newly sequenced mitochondrial genomes (i.e., *H. menziesii* and *C. monodonta*; Guerra et al., in prep). Newly sequenced genomes were sequenced at the sequencing platform of McGill University [Montreal, Canada] using the genome sequencer FLX sequencing service, and all others were obtained by Sanger sequencing [40, 42]. All species and GenBank entries used in

this study are listed in Table I (Note: M-*orf* sequences for *Lasmigona complanata, Margaritifera margaritifera and Toxolasma lividus* have not been obtained). The sequences were translated with ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/; [49]) using the invertebrate mitochondrial genetic code, and analyzed at the nucleotide and/or amino acid level (see below). Because M-*ORF* and F-*ORF* protein sequences vary little within a species, only one sequence was used for each gonochoric species. H-*ORF* sequences are highly variable within species [42], and so multiple sequences were analyzed per species to provide a more complete picture of intraspecific H-*ORF* evolution and potential functionality.

Table I. Sequences analyzed in the present study

Species	**mtDNA type**	**Accession number**	**ORF name**
Subfamiliy Ambleminae			
Quadrula quadrula	M	FJ809751.1	Qqu-M*orf*
	M	FJ809751.1	Qqu-M*cox1*
	F	FJ809750.1	Qqu-F*orf*
	F	FJ809750.1	Qqu-F*cox1*
Toxolasma lividus	F	HM849457.1	Tli-F*orf*
Toxolasma parvum	H	To come	Tpa-H*orf*
Venustaconcha ellipsiformis	M	FJ809752.1	Vel-M*orf*
	M	FJ809752.1	Vel-M*cox1*
	F	FJ809753.1	Vel-F*orf*
	F	FJ809753.1	Vel-F*cox1*
Subfamiliy Anodontinae			
Anodonta anatina	M	KF030962.1	Aan-M*orf*
	F	KF030964.1	Aan-F*orf*

Subfamiliy Gonideinae			
Inversidens japanensis	M	AB055624.1	Ija-M*orf*
	M	AB055624.1	Ija-M*cox1*
	F	AB055625.1	Ija-F*orf*
	F	AB055625.1	Ija-F*cox1*
Solenaia carinatus	M	KC848655.1	Sca-M*orf*
	M	KC848655.1	Sca-M*cox1*
	F	KC848654.1	Sca-F*orf*
	F	KC848654.1	Sca-F*cox1*
Subfamiliy Hyriidae			
Hyridella menziesii	M	Guerra et al. in prep	Hme-M*orf*
	M	To come	Hme-M*cox1*
	F	To come	Hme-F*orf*
	F	AY785394.1	Hme-F*cox1*
Subfamiliy Margaritiferinae			
Cumberlandia monodonta	M	Guerra et al. in prep	Cmo-M*orf*
	M	To come	Cmo-M*cox1*
	F	HM849375.1	Cmo-F*orf*
	F	KF647374.1	Cmo-F*cox1*
Margaritifera falcata	H	HM849545.1	Mfa-H*orf* (1-4)
	H	HM856634.1	
	H	HM849547.1	
	H	HM849548.1	
	H	HM856634.1	Mfa-H*cox1*(1-2)
	H	NC_015476.1	
Margaritifera margaritifera	F	HM849399.1	Mma-F*orf*
	F	HM849095.1	Mma-F*cox1*

Subfamiliy Unioninae			
Lasmigona complanata	F	HM849393.1	Lco-F*orf*
Lasmigona compressa	H	HM849534.1	Lco-H*orf* (1-2)
	H	HM849535.1	
	H	HM856638.1	Lco-H*cox1*(1-2)
	H	NC_015481.1	
Lasmigona subviridis	H	HM849542.1	Lsu-H*orf* (1-2)
	H	HM849543.1	
Pyganodon grandis	M	FJ809755.1	Pgr-M*orf*
	M	FJ809755.1	Pgr-M*cox1*
	F	FJ809754.1	Pgr-F*orf*
	F	FJ809754.1	Pgr-F*cox1*
Utterbackia imbecillis	H	HM849591.1	
	H	HM849595.1	Uim-H*orf* (1-7)
	H	HM849594.1	
	H	HM849601.1	
	H	HM849606.1	
	H	HM849597.1	
	H	HM849584.1	
	H	NC_015479	Uim-H*cox1*(1-2)
	H	HM856637.1	
Utterbackia peninsularis	M	HM856635.1	Upe-M*orf*
	M	HM856635.1	Upe-M*cox1*
	F	HM856636.1	Upe-F*orf*
	F	HM856636.1	Upe-F*cox1*

NOTE – M = M mtDNA in a DUI gonochoric breeding system, F = F mtDNA in a DUI gonochoric breeding system, H = H mtDNA in a non-DUI hermaphroditic breeding system.

Section 2.2.2: Analyses of ORFan sequences and protein secondary structures

Alignments of ORFan and *COX1* sequences were performed with M-COFFEE (DNA) and PSI-COFFEE (proteins) (http://tcoffee.crg.cat/apps/tcoffee/index.html; [50]). Nucleotide and amino acid p-distances, as well as a codon-based test of positive selection using the Nei-Gojobori method [51] were calculated using MEGA6 [52] with variance estimated using 500 bootstrap repetitions. The program VISTA [53] was also used to display the level of sequence conservation between M vs. M, F vs. F, and F vs. H complete mitochondrial genomes. M- and F-type mtDNAs were not compared due to their high divergence and previous characterization [40]. Hydropathy profiles of each amino acid sequence were calculated with the ProtScale tool at ExPASy (http://ca.expasy.org/tools/; [54]) using the method of Kyte and Doolittle [55]. Putative transmembrane (TM) helices were identified using a variety of protein signature recognition methods implemented by the following programs: Phobius (http://phobius.sbc.su.se/; [56]), InterProScan (TMHMM) (http://www.ebi.ac.uk/Tools/pfa/iprscan5/; [57]), TMPred (http://www.ch.embnet.org/software/TMPRED_form.html; [58]), TOPCONS (http://topcons.cbr.su.se/; [59]), and Predict Protein (http://www.predictprotein.org; [60]).

Section 2.2.3: Functional analyses of ORFan proteins

Signal peptides (SPs) were sought using Phobius, InterProScan, PrediSi (http://www.predisi.de/; [61]), and SignalP (http://www.cbs.dtu.dk/services/SignalP/; [62]). Motif Scan (http://myhits.isb-sib.ch/cgi-bin/motif_scan; [63]) and HHpred (http://toolkit.tuebingen.mpg.de/hhpred; [64]) were used to search for motifs and functional domains. TPRpred (http://toolkit.tuebingen.mpg.de/tprpred; [65]) was used to search for potential tetratricopeptide repeat (TPR) or pentatricopeptide repeat (PPR) motifs. The following procedure was used to predict the function of ORFan proteins: (1) we performed BLASTp, tBLASTx, and PSI-BLAST searches against NCBI's entire non-redundant protein

database (NRDB) and against mitochondrial proteins only (last accessed July, 2015) with default parameters (http://blast.ncbi.nlm.nih.gov/; [66]); (2) we used hmmbuild (v3.1b2; downloaded from http://hmmer.janelia.org) [67] to generate two HMM profiles from both the F-*ORF* and M-*ORF* protein alignments (H-*ORF*s were not considered given their scattered phylogenetic distribution and independent evolutionary histories) using default and custom parameters (four profiles in total), and performed profile HMM – sequence comparisons against UniProtKB, Swissprot, PDB, QfO, and Pfamseq databases using HMMER hmmsearch (http://www.ebi.ac.uk/Tools/hmmer/; [67]) with default parameters (E-value cutoff = 0.001); (3) for profile HMM – profile HMM comparisons, we used HHpred, which compares HMM profiles with databases of HMMs representing proteins with known structure (e.g. PDB, SCOP) or annotated protein families (e.g. PFAM, SMART, CDD, COGs, KOGs); and (4) the following programs were also used to predict the function of ORFan proteins: @tome2 (http://atome.cbs.cnrs.fr/; [68]), I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/; [69]) and PredictProtein, which also returned the predicted subcellular localization, binding sites, and ligands (with no accompanying measure of significance). For BLASTp and PredictProtein all matches with E-values <1.0 were kept, while for position-specific iterative or PSI-BLAST all matches with E-values <0.01 were kept as recommended by the program (except for PSI-BLAST analyses against NCBI mitochondrial genes only, where E-values <1.0 were kept, see below). For I-TASSER, top templates with Z scores over 1.0 and structural analogs with TM scores over 0.5 were recorded. All @tome2 results were kept. Motif Scan results not marked as "questionable" or "weak" were kept. Hits described as "uncharacterized," "putative," "unknown," or "predicted" were not kept. Hits from all programs were scored in 122 categories covering putative ligands, subcellular localization, and protein functions, and analyzed.

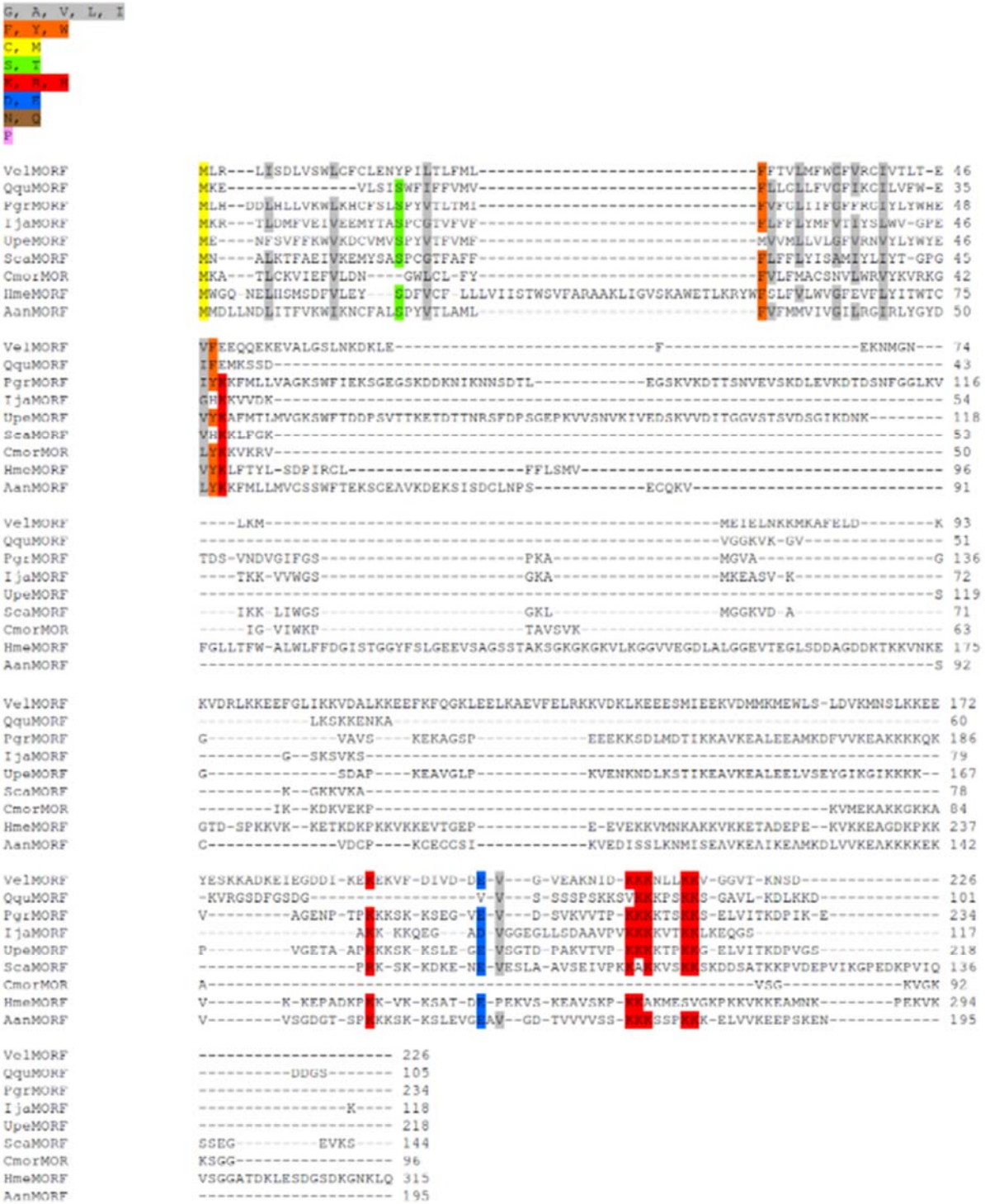
Section 2.3: Results

Section 2.3.1: Rate of evolution of ORFan genes and proteins

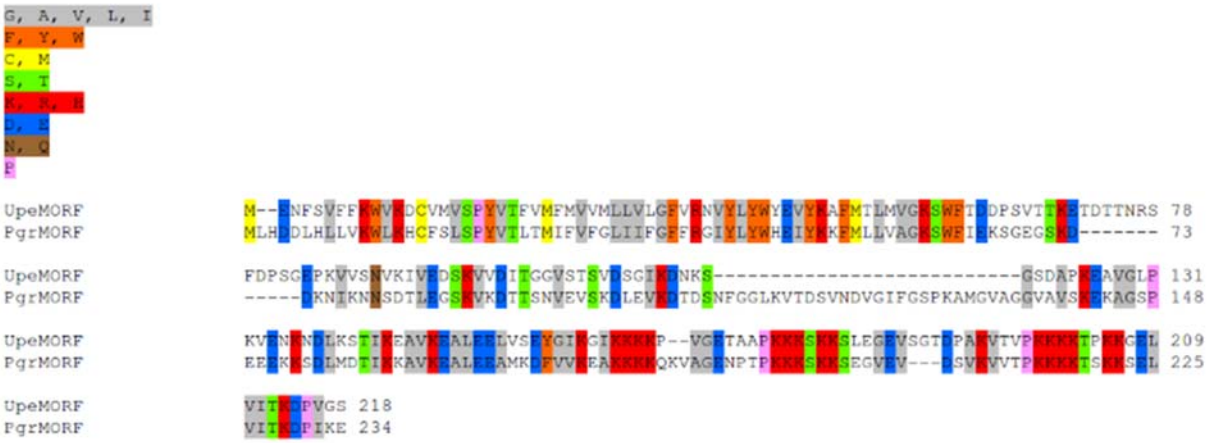
The amino acid sequences of ORFans were generally not well conserved among unionoid species. As seen in Figure 3, a good comprehensive alignment including all M-*ORF* sequences was not possible due to their high divergence, however, sequences from the same

subfamily produced good alignments: *P. grandis* and *U. peninsularis* (Unionidae: Unioninae); *I. japanensis* and *S. carinatus* (Unionidae: Gonideinae); *Q. quadrula* and *V. ellipsiformis* (Unionidae: Ambleminae) (Figure 3b-d). A common feature of M-*ORFs* is that they are all lysine-rich proteins with poly-K strings found in many of them, a characteristic that is apparently absent in F-*ORF* and H-*ORF* amino acid sequences. Similar to M-*ORF* sequences, F-*ORF* sequences from the same subfamily or family, produced better alignments than for all species: *L. complanata, P. grandis* and *U. peninsularis* (Unionidae: Unioninae); *I. japanensis* and *S. carinatus* (Unionidae: Gonideinae); *Q. quadrula, T. lividus* and *V. ellipsiformis* (Unionidae: Ambleminae); *C. monodonta* and *M. margaritifera,* (Margaritiferidae: Margaritiferinae) (Figures 3e-i). Finally, because the H-*ORFs* were likely formed by 5 independent evolutionary events [42], these gene sequences are only conserved within a species; interspecific alignment is therefore not possible for hermaphrodites, and alignments between hermaphrodite H-*ORFs* and closely-related gonochoric species F-*ORFs* were mainly of low quality (Supplementary Figure 2).

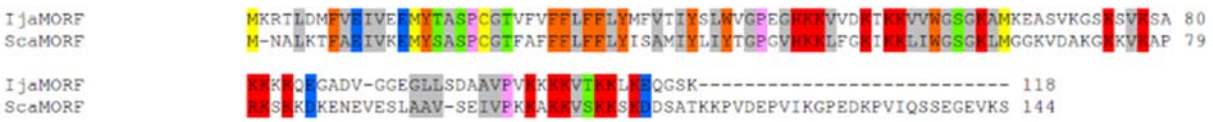
a)


G, A, V, L, I
F, Y, W
C, M
S, T
K, R, H
D, E
N, Q
P
VelMORF
QquMORF
PgrMORF
IjaMORF
UpeMORF
ScaMORF
CmorMOR
HmeMORF
AanMORF

b)


G, A, V, L, I
F, Y, W
C, M
S, T
K, R, H
D, E
N, Q
P
UpeMORF
PgrMORF
UpeMORF
PgrMORF
UpeMORF
PgrMORF
UpeMORF
PgrMORF

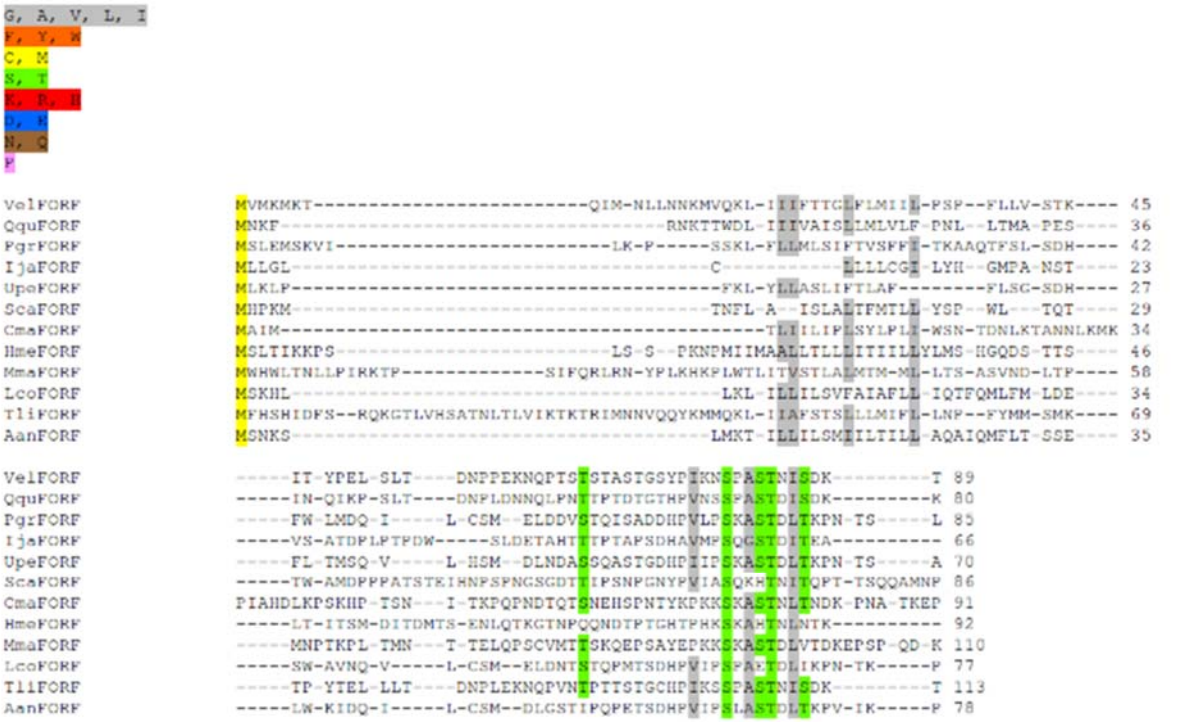
c)


IjaMORF
ScaMORF
IjaMORF
ScaMORF

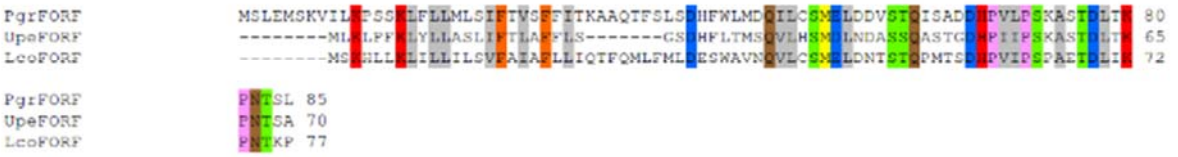
d)


VelMORF
QquMORF
VelMORF
QquMORF
VelMORF
QquMORF

e)


G, A, V, L, I
F, Y, W
C, M
S, T
K, R, H
D, E
N, Q
P
VelFORF
QquFORF
PgrFORF
IjaFORF
UpeFORF
ScaFORF
CmaFORF
HmeFORF
MmaFORF
LcoFORF
TliFORF
AanFORF

f)


PgrFORF
UpeFORF
LcoFORF

g)


ScaFORF
IjaFORF

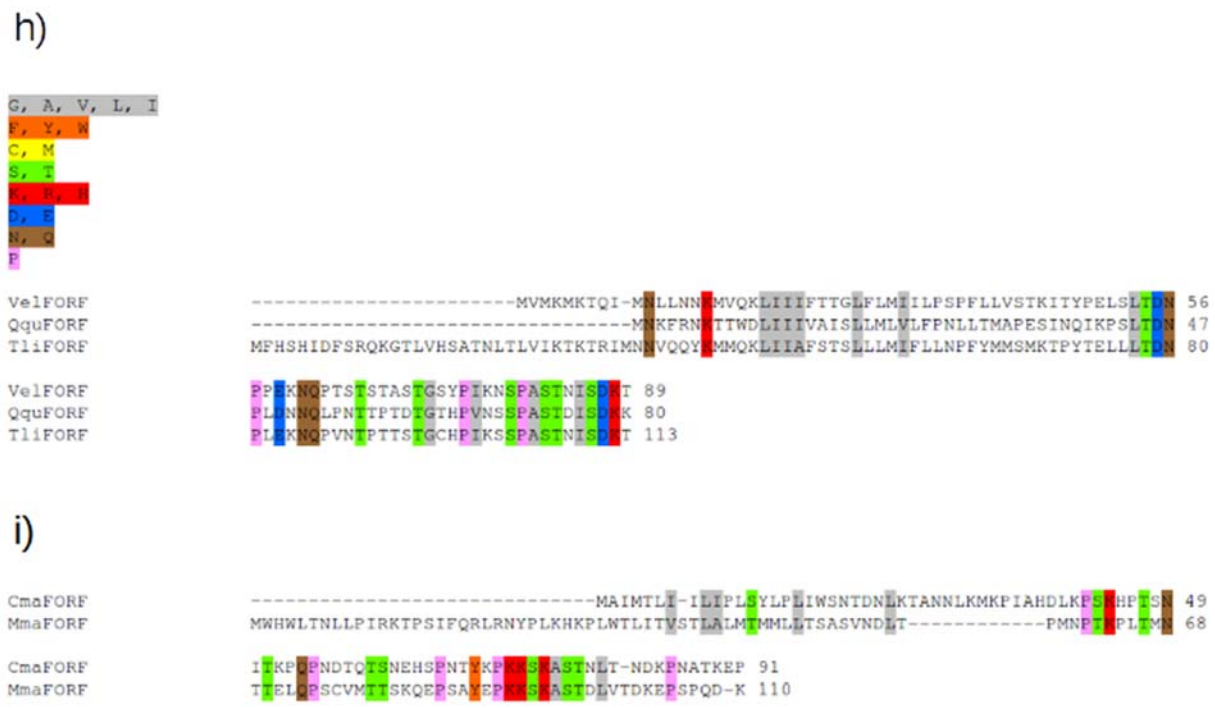


Figure 4. Alignment of M-*ORF* and F-*ORF* protein sequences. Global alignments and alignments for each subfamily are shown. a) All M-*ORF* sequences, b) M-*ORFs* from the subfamily Unioninae, c) M-*ORFs* from the subfamily Gonideinae, d) M-*ORFs* from the subfamily Ambleminae, e) all F-*ORF* sequences, f) F-*ORF* sequences from the subfamily Unioninae, g) F-*ORF* sequences from the subfamily Gonideinae, h) F-*ORF* sequences from the subfamily Ambleminae i) F-*ORF* sequences from the subfamily Margaritiferidae. Colour coding is applied to amino acid groups conserved in ≥70% of sequences. Grey, aliphatic amino acids; orange, aromatic amino acids; yellow, sulfur amino acids; green, amino acids bearing a hydroxyl group; red, basic amino acids; blue, acidic amino acids; brown, amino acids with an amide group; pink, cyclic amino acids. VelMORF, *V. ellipsiformis* M-*ORF*; QquMORF, *Q. quadrula* M-*ORF*; PgrMORF, *P. grandis* M-*ORF*; IjaMORF, *I. japanensis* M-*ORF*; UpeMORF, *U. peninsularis* M-*ORF*; ScaMORF, *S. carinatus* M-*ORF*; CmoMORF, *C. monodonta* M-*ORF*; HmeMORF, *H. menziesii* M-*ORF*; AanMORF, *A. anatina* M-*ORF*; VelFORF, *V. ellipsiformis* F-*ORF*; QquFORF, *Q. quadrula* F-*ORF*; PgrFORF, *P. grandis* F-*ORF*; IjaFORF, *I. japanensis* F-*ORF*; UpeFORF, *U. peninsularis* F-*ORF*; ScaFORF, *S. carinatus* F-*ORF*; CmoFORF, *C. monodonta* F-*ORF*; HmeFORF, *H. menziesii* F-*ORF*;

MmaFORF, *M. margaritifera* F-*ORF*; LcoFORF, *L. complanata* F-*ORF*; TliFORF, *T. lividus* F-*ORF* AanFORF, *A. anatina* F-*ORF*.

The p-distances for nucleotide and amino acid ORFan sequences as well as the outcome of the test of positive selection are reported in Table II (M-*ORFs* and F-*ORFs*) and Table III (H-*ORFs*), along with the values for *cox1* sequences taken from the same sex-specific mtDNAs. Table IV shows the p-distances for within-genus comparisons of F-*ORFs* versus H-*ORFs*. In all cases, the novel ORFs have interspecific p-distances several times higher than *cox1*, which is typically the slowest-evolving protein-coding gene in animal mtDNAs [70]. All groups of sequences also returned a 100% chance of rejecting the null hypothesis of neutral selection in favor of the alternative hypothesis of positive selection. The level of sequence conservation between M vs. M, F vs. F, and F vs. H complete mitochondrial genomes also confirmed that mitochondrial ORFans are the fastest evolving genes in the mtDNA of freshwater mussels with DUI (Supplementary Figure 2).

Table II. p-distances (p-D) and standard error (SE) values for mitochondrial M-*orfs*, F-*orfs*, and *cox1* in freshwater mussel subfamilies

Subfamily	Gene (N)	Nucleotide		Amino acid		p
		p-D	SE	p-D	SE	
Unioninae	F-*orf* (3)	0.355	0.023	0.467	0.047	1.000
	F-*cox1* (2)	0.103	0.007	0.014	0.005	
	M-*orf* (2)	0.350	0.018	0.502	0.034	1.000
	M-*cox1* (2)	0.165	0.010	0.094	0.012	
Gonideinae	F-*orf* (2)	0.469	0.033	0.692	0.058	1.000
	F-*cox1* (2)	0.132	0.008	0.033	0.008	
	M-*orf* (2)	0.384	0.025	0.552	0.044	1.000
	M-*cox1* (2)	0.175	0.009	0.130	0.015	

Ambleminae	F-*orf* (3)	0.351	0.024	0.508	0.041	1.000
	F-*cox1* (2)	0.128	0.009	0.033	0.007	
	M-*orf* (2)	0.421	0.027	0.687	0.047	1.000
	M-*cox1* (2)	0.179	0.010	0.145	0.015	
Margaritiferinae	F-*orf* (2)	0.393	0.029	0.705	0.050	1.000

NOTE – N = number of sequences used. The probability of rejecting the null hypothesis of strict-neutrality ($d_N = d_S$) in favor of the alternative hypothesis ($d_N > d_S$) (in the p column) is shown. d_S and d_N are the numbers of synonymous and nonsynonymous substitutions per site, respectively.

Table III. p-distances (p-D) and standard error (SE) values of mitochondrial H-*orf*s and *cox1* in hermaphroditic freshwater mussels

Species	Gene (N)	Nucleotide		Amino acid		p
		p-D	SE	p-D	SE	
Utterbackia imbecillis	H-*orf* (7)	0.070	0.008	0.181	0.022	1.000
	cox1 (2)	0.000	0.000	0.000	0.000	
Margaritifera falcata	H-*orf* (4)	0.003	0.002	0.004	0.004	1.000
	cox1 (2)	0.000	0.000	0.000	0.000	
Lasmigona compressa	H-*orf* (2)	0.029	0.007	0.065	0.017	1.000
	cox1 (2)	0.000	0.000	0.000	0.000	
Lasmigona subviridis	H-*orf* (2)	0.016	0.005	0.021	0.010	1.000

NOTE – N = number of sequences used. The probability of rejecting the null hypothesis of strict-neutrality ($d_N = d_S$) in favor of the alternative hypothesis ($d_N > d_S$) (in the p column) is shown. d_S and d_N are the numbers of synonymous and nonsynonymous substitutions per site, respectively.

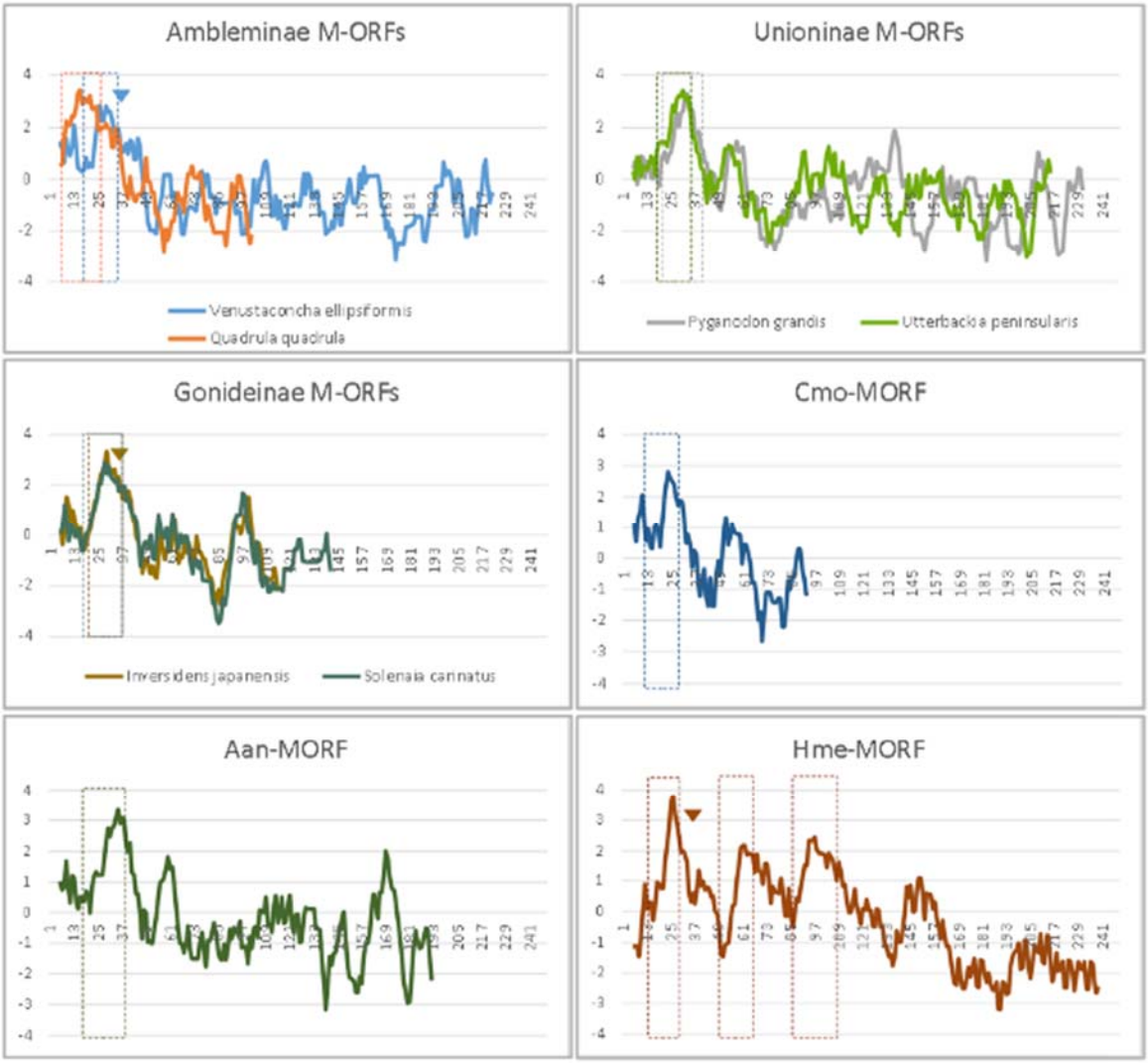
Table IV. p-distances (p-D) and standard error (SE) values of mitochondrial F-*orfs* vs H-*orfs* and F*cox1* vs H*cox1* in comparisons between gonochoric vs. closely related hermaphroditic freshwater mussel species

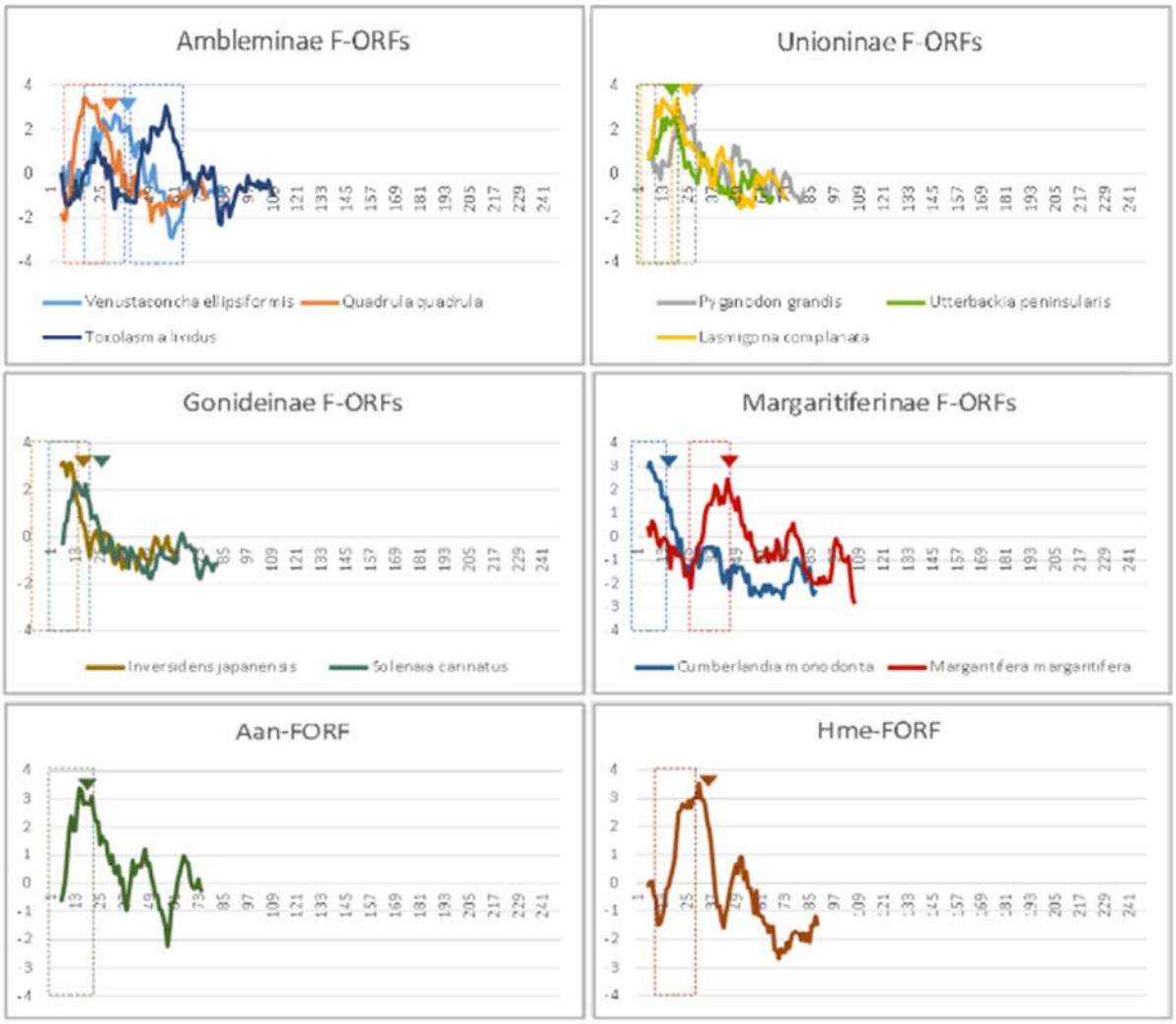
Species	Genes	Nucleotide		Amino acid	
		p-D	SE	p-D	SE
Utterbackia imbecillis vs U. peninsularis	F-*ORF* & H-*ORF1*	0.338	0.034	0.691	0.055
	F-*ORF* & H-*ORF2*	0.310	0.032	0.721	0.054
	F-*ORF* & H-*ORF3*	0.343	0.031	0.743	0.051
	F-*ORF* & H-*ORF4*	0.335	0.034	0.729	0.054
	F-*ORF* & H-*ORF5*	0.333	0.031	0.714	0.052
	F-*ORF* & H-*ORF6*	0.333	0.031	0.714	0.052
	F-*ORF* & H-*ORF7*	0.310	0.030	0.739	0.055
	Mean	**0.329**	**0.005**	**0.722**	**0.007**
	F-*COX1* & H-*COX1-1*	0.547	0.012	0.020	0.006
	F-*COX1* & H-*COX1-2*	0.547	0.012	0.020	0.006
	Mean	**0.547**	**0.000**	**0.020**	**0.000**
Margaritifera falcata vs M. margaritifera	F-*ORF* & H-*ORF1*	0.339	0.025	0.491	0.048
	F-*ORF* & H-*ORF2*	0.336	0.026	0.491	0.049
	F-*ORF* & H-*ORF3*	0.358	0.024	0.491	0.049
	F-*ORF* & H-*ORF4*	0.336	0.026	0.491	0.049
	Mean	**0.342**	**0.005**	**0.491**	**0.000**
	F-*COX1* & H-*COX1-1*	0.469	0.022	0.000	0.000
	F-*COX1* & H-*COX1-2*	0.469	0.021	0.000	0.000
	Mean	**0.469**	**0.000**	**0.000**	**0.000**
Lasmigona complanata vs L. compressa	F-*ORF* & H-*ORF1*	0.218	0.028	0.394	0.059
	F-*ORF* & H-*ORF2*	0.255	0.027	0.395	0.055
	Mean	**0.237**	**0.019**	**0.395**	**0.000**

Lasmigona complanata vs L. subviridis	F-*ORF* & H-*ORF*1	0.269	0.029	0.429	0.054
	F-*ORF* & H-*ORF2*	0.295	0.029	0.442	0.055
	Mean	**0.282**	**0.013**	**0.436**	**0.007**
Toxolasma parvum vs T. lividus	F-*ORF* & H-*ORF*	0.443	0.027	0.736	0.044

Section 2.3.2: Conserved structures in ORFan protein sequences

One TM helix was predicted near the N-terminus of all M-*ORFs* (Figure 4 and Supplementary Table I), except for *H. menziesii* M-*ORF* sequence, for which one N-terminal and two additional TM helices were predicted. PrediSi and SignalP both returned predicted SPs for all M-*ORF* sequences, however, the programs rarely agreed about the length of the predicted signal peptide (Supplementary Table II). One TM helix was also predicted in all F-*ORF* sequences, with a SP predicted to overlap with this TM structure, except in the case of *T. lividus* F-*ORF* where the location of the SP was uncertain (Figure 4 and Supplementary Tables I and II). All H-*ORFs* contained one predicted TM helix near the N-terminus as well, except for *U. imbecillis* H-*ORFs* that contained multiple predicted TM helices, only the first of which had a confident location (Figure 4 and Supplementary Table III). *U. imbecillis* H-*ORFs* also returned variable SP predictions, whereas all other H-*ORF* sequences contain one predicted SP overlapping with the N-terminal TM helix (Supplementary Table IV). Although they could not be confidently aligned (see Supplementary Figure 2), F-*ORFs* and H-*ORFs* of closely related species showed some structural similarities in the localization of the TM helices and SPs (Figure 4). Importantly, all H-*ORFs* contain tandem repeats (*L. compressa* possesses between 3 to 7 tandemly repeated sequence motifs of 20 or 21aa; *L. subviridis* 7 to 9 repeats of 17aa; *T. parvum* 2 to 3 repeats of 47aa; *M. falcata* 2 to 3 repeats of 12aa; and *U. imbecillis* 2 to 4 repeats of 11 or 21aa), which are not found in F-*ORFs* and account for most of the difference in length between F-*ORFs* and H-*ORFs* of closely related species (Supplementary Figure 2).


Ambleminae M-ORFs
Venustaconcha ellipsformis
Quadrula quadrula
Unioninae M-ORFs
Pyganodon grandis
Utterbackia peninsularis
Gonideinae M-ORFs
Inversidens japanensis
Solenaia carinatus
Cmo-MORF
Aan-MORF
Hme-MORF


Ambleminae F-ORFs
Venustaconcha ellipsiformis
Quadrula quadrula
Toxolasma lividus
Unioninae F-ORFs
Pyganodon grandis
Utterbackia peninsularis
Lasmigona complanata
Gonideinae F-ORFs
Inversidens japanensis
Solenaia carinatus
Margaritiferinae F-ORFs
Cumberlandia monodonta
Margaritifera margaritifera
Aan-FORF
Hme-FORF

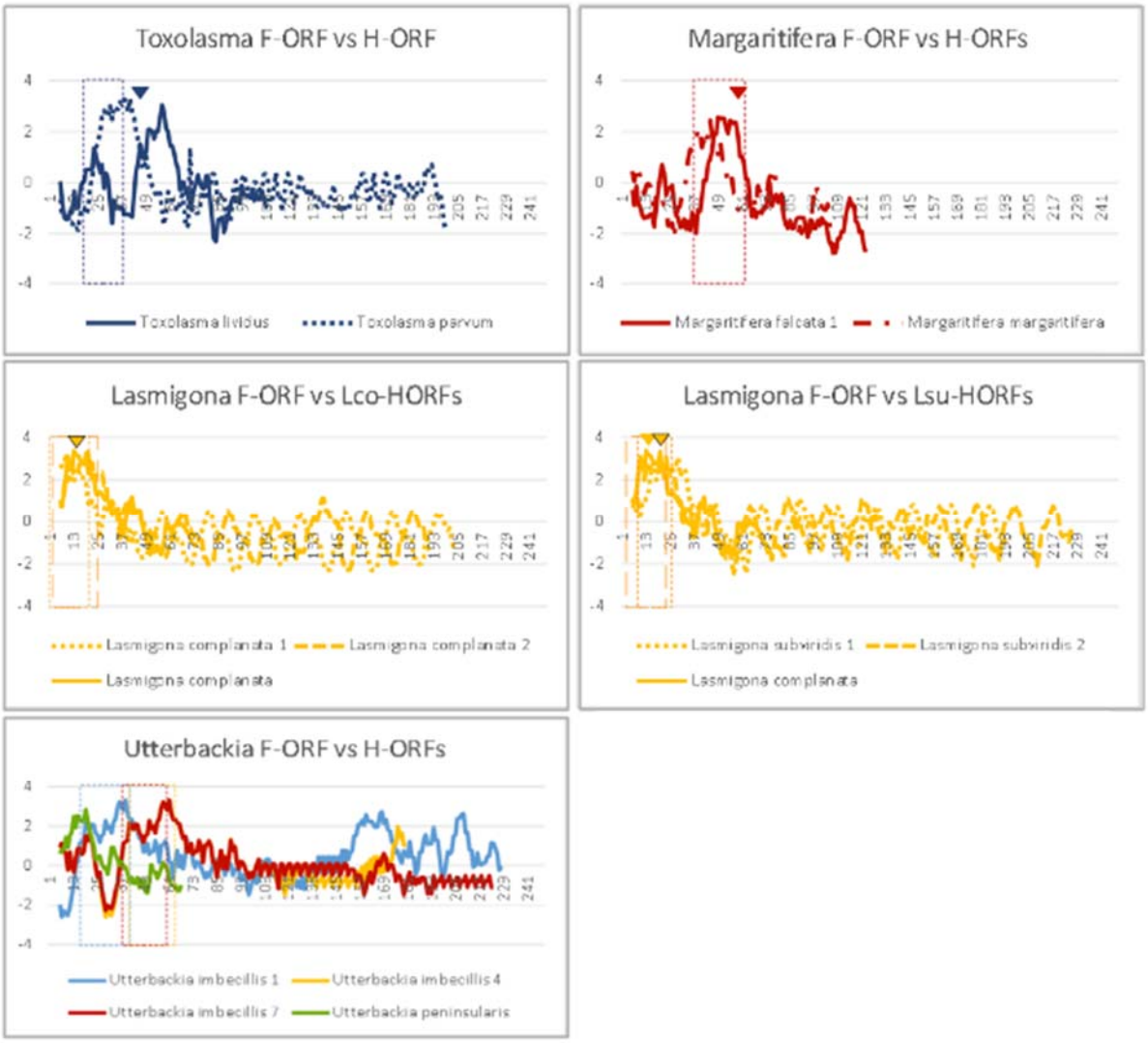


Figure 5. Hydrophobicity profiles of M-*ORFs* (a), F-*ORFs* (b) and H-*ORFs* vs. F-*ORFs* (c). Boxes indicate predicted TM helices, arrowheads indicate the end of predicted SPs. X-axis is amino acid position, Y-axis is hydrophobicity. Margaritifera H-*ORFs*: Mfa1 and Mfa2&4 have nearly identical profiles. Lasmigona H-*ORFs*: arrowheads outlined in black indicate the end of the SP in sequence 1, arrowhead without outline is for Lco-HORF2; boxes with long dashes are for sequence #2. Utterbackia H-*ORFs*: Sequences 2-6 have profiles similar to that of sequence 7.

Section 2.3.3: Motif and functional domain scans: frequently recurring HHpred hits and potential ligand-binding sites

Six HHpred hits consistently appeared highly ranked in the results of M-*ORFs*, F-*ORFs* and H-*ORFs*: (1) prepilin-type processing-associated H-X9-DG domain, (2) outer membrane insertion C-terminal signal, (3) LPXTG cell wall anchor domain, (4) X-X-X-Leu-X-X-Gly heptad repeats, (5) GlyGly-CTERM domain and (6) a pentatricopeptide repeat (PPR) domain. Probabilities were all >92% (which the developers state can be interpreted literally [64]), and ranks were typically 1-6 in variable order, with very few of these hits falling outside of the top 10 (Supplementary Tables V and VI). Figure 5 shows the position of these six hits in the protein sequences analyzed. Other but less recurring motifs and domains are presented in detail in Supplementary Tables VII and VIII. No TPR or PPR motifs were found.

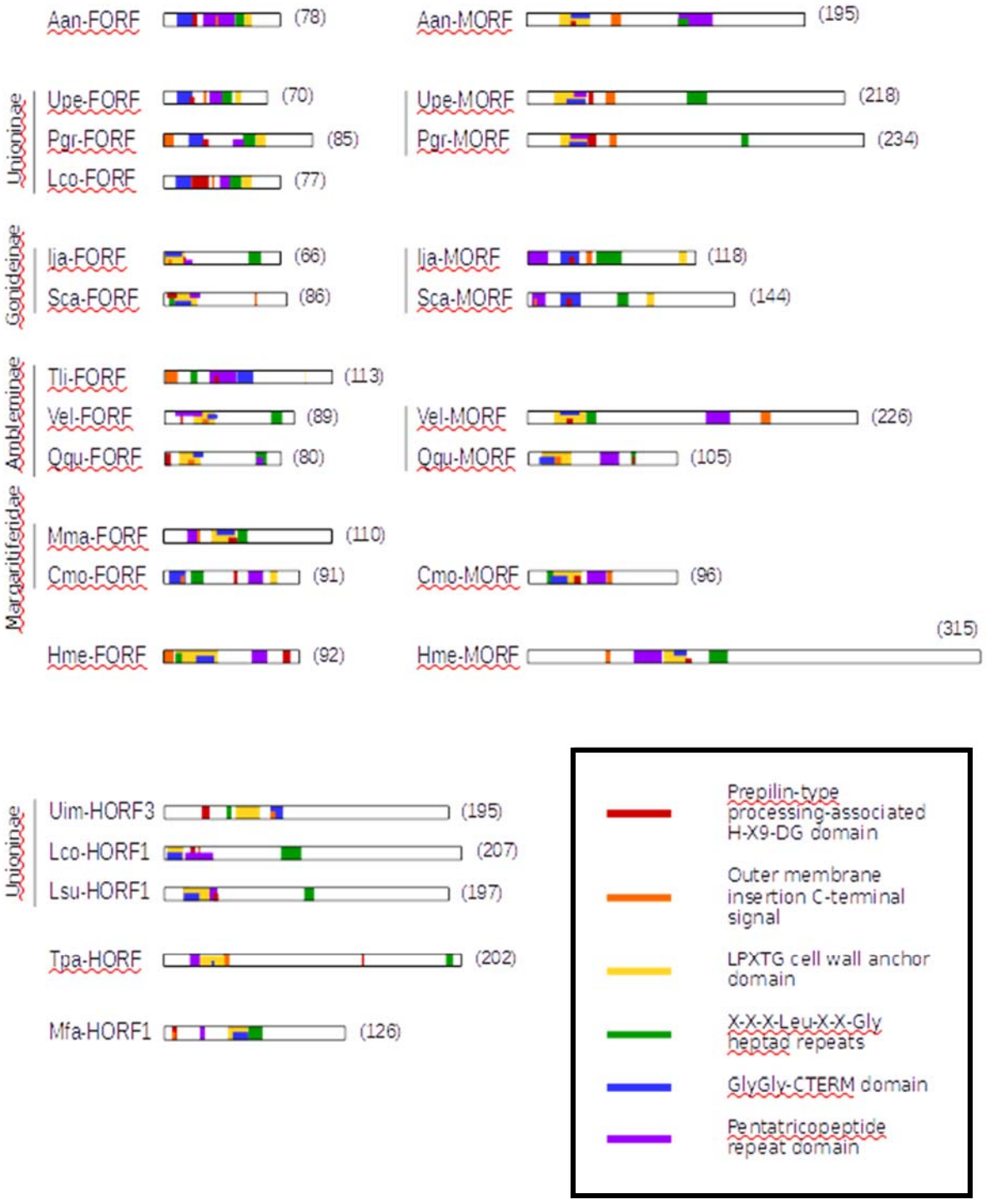


Figure 6. Position of motifs frequently recurring in HHpred hits. Protein length in amino acids is indicated in parentheses. One representative sequence was chosen for each hermaphroditic species.

Inferred homologies and prediction of binding sites both indicated that ORFan proteins may bind several ligands (Table V). All M-*ORFs* returned hits to protein-binding, DNA-binding and RNA-binding proteins, while many returned hits to proteins binding ions (8 species), ATP (6 species), carbohydrates (3 species), and lipids (3 species). All F-*ORFs* returned hits to protein-binding and RNA-binding proteins, while many returned hits to proteins binding DNA (11 species), ions (10 species), ATP (8 species), carbohydrates (7 species), and lipids (4 species). H-*ORF* sequences returned hits to proteins binding other proteins (5 species), DNA (5 species), RNA (5 species), carbohydrates (5 species), ions (4 species), ATP (3 species), and lipids (2 species).

Table V. Summary of hits to ligand-binding sites in M-*ORFs*, F-*ORFs* and H-*ORFs*

Protein	DNA	RNA	Protein	Carbohydrate	Ion	Lipid	ATP
Vel-MORF	X	X	X	X		X	X
Qqu-MORF	X	X	X		X		
Pgr-MORF	X	X	X	X	X		X
Ija-MORF	X	X	X		X		X
Upe-MORF	X	X	X		X		X
Sca-MORF	X	X	X		X		X
Cmo-MORF	X	X	X		X	X	
Hme-MORF	X	X	X		X	X	X
Aan-MORF	X	X	X	X	X		
Total	**9**	**9**	**9**	**3**	**8**	**3**	**6**
Vel-FORF	X	X	X	X	X		X
Qqu-FORF	X	X	X	X	X	X	X
Pgr-FORF	X	X	X	X	X		X
Ija-FORF	X	X	X	X	X		
Upe-FORF	X	X	X		X		X
Sca-FORF	X	X	X		X	X	X
Cmo-FORF		X	X	X			

Hme-FORF	X	X	X	X			X
Lco-FORF	X	X	X		X	X	X
Tli-FORF	X	X	X	X	X		X
Mma-FORF	X	X	X		X		
Aan-FORF	X	X	X		X	X	
Total	**11**	**12**	**12**	**7**	**10**	**4**	**8**
Uim-HORF1 - 3	X	X	X	X	X		X
Uim-HORF4 - 7	X	X	X	X	X		
Mma-HORF1, 2, 4	X	X	X		X		X
Mma-HORF3	X	X	X		X		
Tpa-HORF	X	X	X	X	X	X	X
Lco-HORF1	X	X	X		X		
Lco-HORF2	X	X	X	X	X	X	
Lsu-HORF1 - 2	X	X	X	X	X		
Total	**14**	**14**	**14**	**10**	**14**	**2**	**6**

Section 2.3.4: Prediction of molecular function: hits to viral proteins

Because a viral origin for the mitochondrial ORFans in DUI bivalves has previously been suggested [44], our results obtained with all programs for protein function prediction (i.e., BLAST, HMMER, HHpred, @tome2, I-TASSER, and PredictProtein) were first scanned for supported hits to viral proteins (Table VI). Overall, H-*ORFs* returned more viral hits than M-*ORFs* or F-*ORFs*. *M. falcata* H-*ORFs* primarily returned envelope proteins, *L. subviridis* H-*ORFs* returned capsid and envelope proteins, *L. compressa* H-*ORFs* returned proteins that interact with receptors, and *T. parvum* H-*ORF* returned a protein that regulates the degradation of a receptor. *U. imbecillis* H-*ORFs* returned many copies of capsid proteins and other structural proteins. M-*ORFs* returned nucleoproteins (*A. anatina* and *H. menziesii*), membrane proteins (*I. japanensis* and *S. carinatus*), and proteins with a role in replication, life cycle, and apoptosis (*A. anatina, U. peninsularis, I. japanensis* and *V. ellipsiformis*). F-*ORF* hits were mostly parts of the viral capsid and viral envelope (*S. carinatus, T. lividus* and *M.*

margaritifera), receptors/fibre proteins (*M. margaritifera* and *C. monodonta*), or proteins involved in cell cycle and translation (*P. grandis* and *I. japanensis*).

Table VI. Hits to viral proteins from structural prediction analyses

Gene	Hit	Function	Position
Aan-MORF	Nucleoprotein, *Andes virus* [Atome 2; 41.16] Regulatory protein MNT, *Enterobacteria phage P22* [Atome 2; 21.14]	**Nucleoprotein** **Gene regulation**	NA NA
Upe-MORF	Uncharacterized protein 56B, *Sulfolobus islandicus* [Atome 2; 27.96]	**Transcription repressor**	NA
Pgr-MORF	Matrix protein 1, *Influenza A virus* [Atome 2; 39.16] Helix-destabilizing protein, *Enterobacteria phage T7* [Atome 2; 18.55]	**Matrix protein** **DNA binding protein**	NA NA
Ija-MORF	Nonstructural protein 5A, *Bovine viral diarrhea virus 1-CP7* [Atome 2; 33.37] Functional anti-apoptotic factor vBCL-2 homolog, *Human herpesvirus 8* [Atome 2; 27.14]	**Membrane protein** **Apoptosis**	NA NA
Sca-MORF	Nonstructural protein 5A, *Bovine viral diarrhea virus 1-CP7* [Atome 2; 22.35]	**Membrane protein**	NA
Vel-MORF	Macrophage galactose N-acetyl-galactosamine specific lectin 2 [Hhpred; 93.40] RhUL123, *Macacine herpesvirus 3* [I-TASSER; TM score 0.671] Phosphoprotein, *Measels virus* [Atome 2; 49.33] Tail needle protein gp26, *Enterobacteria phage P22* [Atome 2; 48.96]	**C-type lectin** **Viral life cycle** **Unknown function** **Fibrous protein**	20-171 NA NA NA
Qqu-MORF	Virion RNA polymerase, *Bacteriophage n4* [I-TASSER; TM score 0.542]	**Transferase**	NA
Cmo-MORF	**No hits to viral proteins**		
Hme-MORF	Nucleoprotein, *Andes virus* [Atome 2; 63.91]	**Nucleoprotein**	NA
Aan-FORF	**No hits to viral proteins**		
Upe-FORF	BM2 protein, *Influenza B virus (B/Taiwan/70061/2006)* [Atome 2; 42.29]	**Transport protein**	NA
Pgr-FORF	V-cyclin, *Human herpesvirus 8* [I-TASSER; norm. TM score 0.517]	**Cell cycle**	NA

Lco-FORF	Herpes simplex virus protein ICP47, *Herpes simplex virus (type 1 / strain 17)* [Atome 2; 46.61]	**Membrane protein**	NA
Ija-FORF	Non-structural RNA-binding protein 34, *Simian rotavirus A/SA11* (2) [Atome 2; 48.04, 28.60]	**Translation**	NA
Sca-FORF	Major capsid protein (protein P3), *Enterobacteria phage PRD1* [Atome 2; 80.01]	**Capsid protein**	NA
Tli-FORF	Envelope protein E, *Dengue virus 2 Thailand/16681/84* [Atome 2; 46.45]	**Envelope protein**	NA
Vel-FORF	V1V2 region of HIV-1 on 1FD6 scaffold, *Human immunodeficiency virus 1* [Atome 2; 57.65]	**Immune system**	NA
Qqu-FORF	HIV-1 matrix protein, *Human immunodeficiency virus 1* (2) [Atome 2; 83.13, 72.79]	**Matrix protein**	NA
Mma-FORF	ODV-E18: Occlusion-derived virus envelope protein ODV-E18 (2) [Hhpred; 72.05, 62.79]	**Envelope protein**	21-62
	Adenovirus fibre, *Human adenovirus 2* [Atome 2; 27.29]	**Fibre protein**	23-55
	Fibre protein 2 (receptor-binding domain), *Human adenovirus 41* [I-TASSER; 18.06]	**Fibre protein, receptor binding**	NA NA
Cmo-FORF	Virus attachment protein globular domain (49835) SCOP seed sequence: d1h7za [Hhpred; 21.78]	**Viral attachment, entry into host cell**	50-68
	Adenovirus fibre protein; cell receptor recognition, receptor, *Human adenovirus type 3* [Hhpred; 21.71]	**Fibre protein, Cell receptor recognition**	44-68
	Fibre protein, *Human adenovirus 37* [Atome 2; 31.21]		NA
	Fibre protein, *Human adenovirus 2* [Atome 2; 30.90]		NA
	Type 5 fibre protein, *Human adenovirus 5* [Atome 2; 30.46]		NA
	Fibre protein, *Human adenovirus 41* [Atome 2; 24.60]		NA
Hme-FORF	Nucleoprotein, *Influenza A virus* [Atome 2; 80.49]	**RNA binding protein**	NA
Uim-HORFs	HIV-1 capsid, *Human immunodeficiency virus 1* [I-TASSER; TM score 0.513]	**Capsid protein**	NA
	Gag Polyprotein, *Human immunodeficiency virus 1* [I-TASSER; TM score 0.510]	**Precursor protein**	NA
	Capsid protein P24, *Human immunodeficiency virus type 2* [I-TASSER; TM score 0.504]	**Capsid protein**	NA
	Nucleoprotein, *Andes virus* [Atome 2; 44.18]	**Nucleoprotein**	NA
	Protein ICP47, *Herpes simplex virus* [Atome 2; 37.48]	**Membrane protein**	NA
	LdOrf-129 peptide, *Lymantria dispar multiple nucleopolyhedrovirus* (2) [BLASTP, PSIBLAST; 2e-06, 7e-10]	**Structual protein** **Unknown**	74-144 74-131
	ORF-132 protein, *Lymantria dispar multiple nucleopolyhedrovirus* (2) [BLASTP, PSIBLAST; 4e-06, 2e-09]	**Unknown**	72-140
	orf-126 protein, *Lymantria dispar multiple nucleopolyhedrovirus* [PSIBLAST; 4e-08]	**Unknown**	60-154
	Central variable region protein, *African swine fever virus* [PSIBLAST; 6e-08, 7e-07]	**Unknown**	60-130

	Central variable region protein, *African swine fever virus* [PSIBLAST; 7e-08]	**Structural capsid protein, chaperone in capsid assembly (several hits)**	65-153
	pB602L, *African swine fever virus tick/South Africa/Pretoriuskop Pr4/1996* [PSIBLAST; 8e-08]		
		Spliceosomal RNA	65-137
		DNA transfer protein	NA
	U1, *Hyposoter didymator ichnovirus* [PSIBLAST; 3e-07]		88-166
	gp7, *Salmonella phage epsilon15* [I-TASSER; norm. Z-score 1.32]	**Fibre protein**	NA
	Long tail fibre protein p37, *Enterobacteria phage T4* [I-TASSER; norm. Z-score 1.30]	**Viral life cycle**	NA
	RhUL123, *Macacine herpesvirus 3* [I-TASSER; TM score 0.617]	**Nucleoporin (several hits)**	NA
	Nucleoprotein, *Andes virus* [Atome 2; 39.59]	**Structual protein**	NA
		Unknown	
	LdOrf-129 peptide, *Lymantria dispar multiple nucleopolyhedrovirus* [PSIBLAST; 8e-10]		87-188
	ORF-132 protein, *Lymantria dispar multiple nucleopolyhedrovirus* [PSIBLAST; 5e-09]	**DNA binding & stabilization**	139-223
	DNA stabilization protein, *Salmonella phage HK620* [I-TASSER; Z-score 1.09]	**Major coat protein**	NA
	Hexon protein, *Human adenovirus 5* [I-TASSER; Z-score 1.01]	**Matrix protein**	NA
	Human T-cell leukemia virus type II matrix protein, *Human T-lymphotropic virus 2* [I-TASSER; Z-score 1.00]	**Blocks the major histocompatibility complex class I antigen presentation pathway**	
	Herpes simplex virus protein ICP47, *Herpes simplex virus (type 1 / strain 17)* [Atome 2; 1.72]		
Lco-HORFs	Long tail fiber protein P37, *Enterobacteria phage T4* [I-TASSER; Z-score 1.01]	**Receptor binding**	NA
	Capsid protein, *Rubella virus strain M33* [Atome 2; 83.05]	**Capsid component**	NA
	VPU protein, *Human immunodeficiency virus 1* [Atome 2; 43.79]	**Regulates degradation of CD4 (several hits)**	NA
Lsu-HORFs	Major capsid protein, *Synechococcus phage Syn5* [I-TASSER; Z-score 1.66]	**Capsid component**	NA
	RhUL123, *Macacine herpesvirus 3* [I-TASSER; TM score 0.547]	**Viral life cycle**	69-195
	Herpes virus major outer envelope glycoprotein (BLLF1) [BLASTP/PSIBLAST; 2.73e-03]	**Envelope protein**	NA
	Short tail fiber protein, *Enterobacteria phage T4* [I-TASSER; Z-score 2.14]	**Structural protein**	NA
	Major capsid protein, *Synechococcus phage Syn5* [I-TASSER; Z-score 2.19]	**Capsid component (several hits)**	NA
	Coat protein, *Enterobacteria phage P22* [I-TASSER; TM score 0.520]	**Coat component**	NA
	Herpes virus major outer envelope glycoprotein (BLLF1) [BLASTP/PSIBLAST; 4.85e-04]	**Envelope protein**	NA

Tpa-HORF	VPU protein (Trans-membrane domain), *Human immunodeficiency virus 1* [Atome 2; 33.16]	**Regulates degradation of receptor molecule CD4 (several hits)**	NA
Mfa-HORFs	ODV-E18: Occlusion-derived virus envelope protein ODV-E18 [Hhpred; 74.97]	**Envelope protein (several hits)**	33-73
	Herpes_TK_C: Thymidine kinase from Herpesvirus C-terminal, *Herpesvirus* (2) [Hhpred; 48.70, 48.13]	**ATP binding, thymidine kinase (several hits)**	33-73
	Adenovirus fibre, *Human adenovirus 2* [Atome 2; 34.11]	**Fibre protein, receptor binding (several hits)**	NA

NOTE – A norm. Z-score>1 indicates a good alignment; a TM-score>0.5 indicates a similar fold with query (I-TASSER, [69]); position = amino acid position in the query sequence; NA = not applicable

Section 2.3.5: Prediction of molecular function: hits to mitochondrial proteins

Besides viral hits, most of the sequences analyzed also returned hits to proteins involved in energy production, including proteins of the mitochondrial electron transport system, so we tested the similarity of the ORFan proteins to standard mtDNA-encoded ones with BLAST searches. Our analyses predicted M-*ORFs* mostly as subunit 5 of the NADH-Ubiquinone Oxidoreductase complex I of the mitochondrial electron transport chain (*NAD5*) for 5 species out of 9, and/or *ATP8* of the ATP synthase complex V for 5 species, but only with very low support (i.e., E-values ranged between 6e-04 and <1.0, the limit chosen for this analysis) (see Table VII). This latter result was also supported by a moderately significant domain hit identified in *C. monodonta* (i.e., pfam02326; plant ATP synthase F0; this family corresponds to subunit 8 of the F0 complex of plants; E-value 4e-03) (see below). For F-*ORFs*, the most recurring hit (8 species out of 12) was subunit 2 of the mitochondrial complex I (*NAD2*), again with quite low support (E-values ranged between 6e-08 and <1.0), whereas BLAST searches of H-*ORFs* principally identified F-*ORFs* (3 species out of 5), with moderate E-values (Table VII).

Table VII. List of BLAST hits for mitochondrial ORFans in freshwater mussels searched against NCBI NRDB mitochondrial proteins

Species Name	**M-*ORFs***	**F-*ORFs***	**H-*ORFs***
Anodonta anatina	NAD7 (0.61) ---	--- *atp9 (0.19)*	
Cumberlandia monodonta	**ATP8 (0.81)** **---**	--- ***nad2 (6e-08)***	
Hyridella menziesi	**ATP8 (0.61)** *nad4 (6e-04)*	**NAD2 (0.33)** ***nad2 (0.022)***	
Lasmigona complanata		--- ***nad2 (0.094)***	
Lasmigona compressa			**F-ORF (4e-05)** ***f-orf (2e-05)***
Lasmigona subviridis			**F-ORF (6e-09)** ***f-orf (2e-05)*** *nad1 (0.64)*
Inversidens japanensis	**ATP8 (0.62)** ***nad5 (0.001)*** ***atp8 (0.048)*** *cox1 (0.15)*	--- ***nad2 (0.22)***	
Margaritifera falcata			COX1 (0.94) ---
Margaritifera margaritifera		NAD5 (0.093) **NAD2 (0.23)** ***nad2 (0.15)***	
Pyganodon grandis	**NAD5 (0.046)** *atp9 (0.30)*	--- *cytb (0.13)*	
Quadrula quadrula	**NAD5 (0.026)** **ATP8 (0.070)** *atp9 (0.30)*	NAD5 (0.31) ***nad2 (0.56)***	
Solenaia carinatus	COX1 (0.41) **NAD5 (0.99)**	--- ***nad2 (0.018)***	

	nad5 (0.33)		
Toxolasma lividus		--- ---	
Toxolasma parvum			**F-ORF (0.020)** ---
Utterbackia imbecillis			--- *nad2 (0.061)*
Utterbackia peninsularis	**NAD5 (0.38)** *nad2 (0.31)*	--- *cox1 (0.056)*	
Venustaconcha ellipsiformis	NAD4 (0.19) CYTB (0.21) **ATP8 (0.94)** *nad4 (0.15)*	NAD4 (0.55) ***nad2 (0.14)***	

NOTE – protein name and (E-values <1.0) identified using PSI-BLAST and *tBLASTx* are indicated above in capital letters and below in italics, respectively. Hits to freshwater mussel mitochondrial ORF homologs are not presented, except for the highly divergent H-*ORFs*.

Section 2.3.6: Profile HMM – sequence comparisons for F-*ORFs* and M-*ORFs*

The hmmsearch analyses with HMM profiles for F-*ORF* and M-*ORF* alignments gave different numbers of hits for default vs. custom profiles. In general, the custom profiles were more "stringent" in terms of hit yield among all databases analysed, giving fewer total results than the default ones. Except for one hit for the M-*ORF* profiles, freshwater mussel ORFan sequences were the only significant hits (E-value <0.001) returned for all profiles, and they will not be considered. Therefore, we will describe all the hits other than unionoids ORFans (even those with E-values higher than the cutoff) in terms of functional recurrence. Results are presented in Supplementary Tables IX and X.

Overall, F-*ORF* hits for both profiles are related to membrane association, virus life cycle, and interaction with nucleic acids (Supplementary Tables IX and X). The M-*ORF* default profile frequently returned hits associated with membranes, related to energy

production in bacteria or eukaryotes, transport or movement, or other functions related to membranes (Supplementary Tables IX and X). The Excalibur domain protein, found two times with borderline significance (E-values 0.0011 and 0.0018), also has functions in DNA binding and repair and transcription regulation. Another recurring function is interaction with RNA, sometimes specific to ribosomal functions or biogenesis (pre-rRNA processing, translation initiation, tRNA modification, poly-(A) RNA binding for nuclear import, posttranscriptional expression regulation). Interactions with amino acids and proteins were also common, including protein transport, protein modification, or involvement in cytoskeleton rearrangements. Some hits suggest the possible insertion of DNA from foreign sources, or nuclease activity. Hits to viral delta antigens of hepatitis delta virus are related to viral life cycle (invasion in host cell and nucleus, replication). The M-*ORF* custom profile returned four additional results, all involved in protein and/or membrane interactions.

Section 2.3.7: Prediction of molecular function (all sequences, all programs except hmmsearch)

Finally, we compiled the results obtained for all ORFs with all other programs for protein function prediction (i.e., BLAST, HHpred, @tome2, I-TASSER, and PredictProtein). Figure 6 summarizes the most frequent categories of hits for freshwater mussel mitochondrial ORFans (i.e., those returned for over 75% of sequences per sex) and Supplementary Figure 3 and Supplementary Tables XI-XLV contain detailed hits and recurring functions. Overall, the most common hits for M-*ORFs* were membrane-associated proteins (Figure 6). M-*ORFs* also returned many hits to proteins involved in transport, cellular signalling, and the cell cycle. The most commonly predicted subcellular localizations for M-*ORFs* were membranes (both cellular and organellar) and the endoplasmic reticulum (ER). F-*ORFs* also returned many hits to membrane-associated proteins. These include several trafficking and transport functions such as SNAP receptors, kinases (such as signalling proteins), trans-Golgi transport proteins including sensors, inhibitors, and transporters, and proteins generally involved in cell adhesion. Finally, immune system proteins were common hits including several SNAP receptors (e.g., v-SNARES involved in cytokine secretion). The mitochondria, Golgi, and ER were predicted subcellular localizations (Figure 6). For H-*ORFs*, structural proteins,

particularly collagen and collagen-like proteins were the most common categories, closely followed by transmembrane proteins. Signalling, transport, and transcription factors were common hits as well (Figure 6).

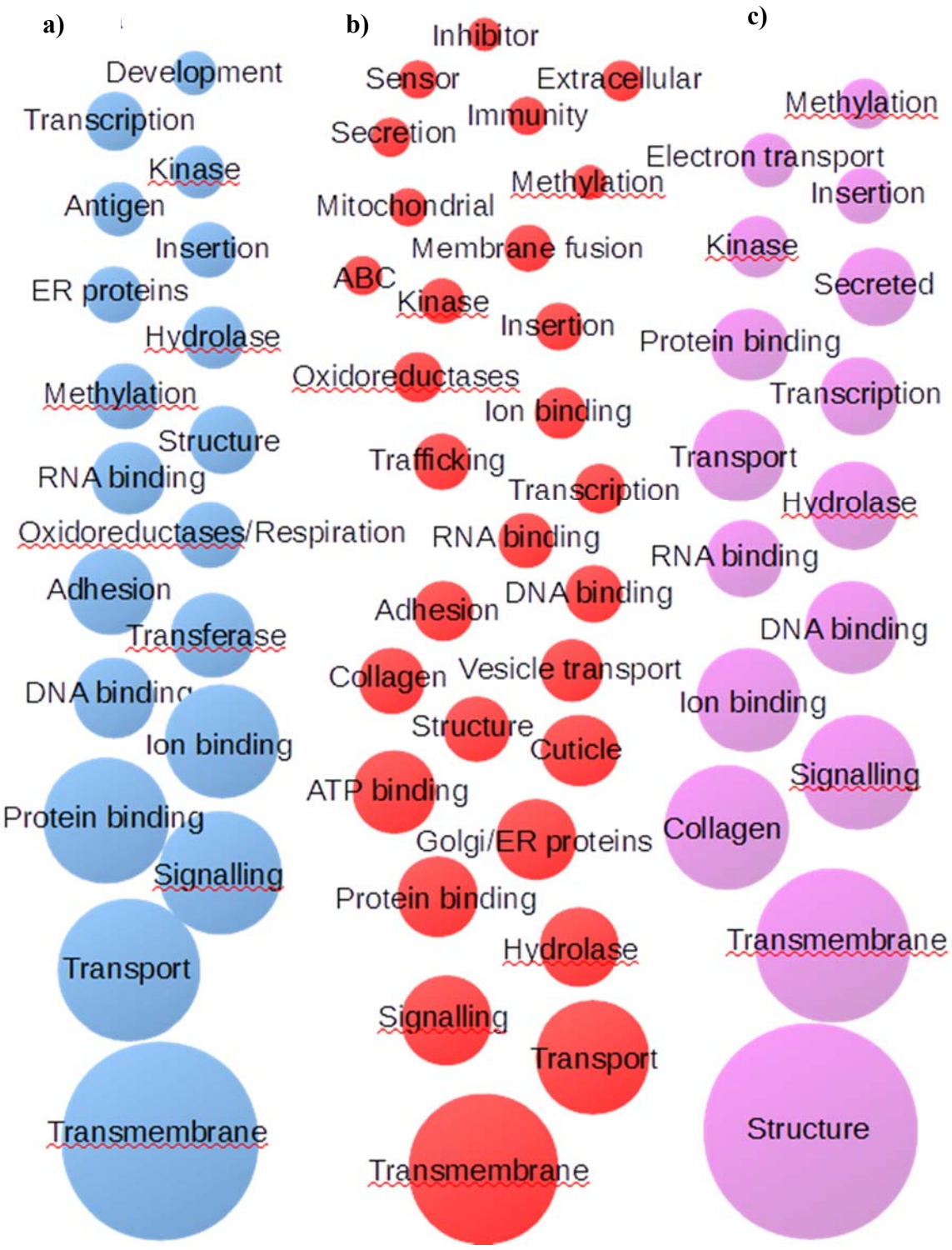


Figure 7. Most common categories of hits for (a) M-*ORFs*, (b) F-*ORFs*, and (c) H-*ORFs*. Bubble size represents average number of hits per sequence analyzed. Largest bubble

(structure, H-*ORFs* = 28.8 hits/sequence), smallest bubble (inhibitor, F-*ORFs* = 0.92 hits/sequence). Blue, M-*ORFs*; red, F-*ORFs*; purple, H-*ORFs*. ER, endoplasmic reticulum; ABC, ATP-binding cassette.

Section 2.4: Discussion and conclusion

Sex-specific mitochondrial DNAs of bivalves with DUI (orders Mytiloida, Veneroida, and Unionoida) harbor ORFans that have previously been hypothesized to have (1) originated from endogenization of viral genes and (2) play a role in the DUI mechanism, e.g. in the maintenance and segregation of sperm mitochondria during male embryo development by masking them from the degradation machinery [44, 46]. However, because the studied taxa (5 mytilids, 1 venerid, and 1 unionid) were evolutionarily distant, and because of the observed structural similarities of the ORFan protein products within evolutionary lineages (i.e., F-*ORF* and M-*ORF* were more similar to each other in the venerid and in the unionid than to the proteins of the same name in the Mytilus complex; in Mytilus spp. F-*ORFs* and M-*ORFs* were respectively more similar among themselves) it was further hypothesized that the ORFans could have originated from independent endogenization events [44, 46]. The evolutionary distance among mytilids, venerids, and unionids did not allow for a good comparison across the species studied [44], and so expanding the number of venerids or unionids used could provide more information on the conservation of these sequences and their characteristics. For these reasons, and because homologous sequences remain more conserved within closely related species (as compared to distantly related ones), we decided to perform *in silico* analyses on more closely related ORFan sequences, i.e., within the order Unionoida, for a better understanding of their putative origin(s) and function(s).

Section 2.4.1: Evolution of freshwater mussel ORFan sequences and protein structures

As previously reported in other DUI bivalves [17], one general feature usually observed in ORFan sequences is their higher p-distance values at the amino acid level compared to their own nucleotide sequences, and the test of positive selection returned a p-value of 1.000 in all cases, rejecting neutral evolution in favor of positive selection. This test

may also indicate relaxed purifying selection – for example all three *ORFs* may be under less purifying selection than the rest of the mitochondrial genome, or *H-ORFs* may be even under less selective pressure than their ancestral *F-ORFs*. This study also directly compared within-genus p-distances of nucleotide and amino acid sequences of M-*ORFs*, F-*ORFs* and H-*ORFs* to a standard mitochondrial gene – *cox1* – for the first time. Among species, unionoid M-*ORF*s display high variability in length and very low extent of amino acid sequence similarity (Figure 1). However, despite limited sequence similarities, M-*ORFs* appear conserved at the secondary structure level, with a single TM helix predicted in the N-terminus half of each M-*ORF*, except for *H. menziesii* (Figure 2). Most M-*ORFs* have an SP predicted in the same region, but given the hydrophobic nature of SPs, many programs struggle to distinguish N-terminal TM helices from SPs, and the number of M-*ORFs* with an SP may actually be higher [56]. Sequence similarities among F-*ORF* sequences are more pronounced, and mostly found within a stretch of ~30-40 residues in the C-terminal region (Figure 1). The relatively conserved region is preceded by a single predicted TM helix as well as one predicted SP in the N-terminus half of each F-*ORF* (Figure 2). Thus, despite low sequence conservation, M-*ORF* and F-*ORF* proteins appear structurally conserved, suggesting that their biological functions might be conserved among species as well.

Contrary to gonochoric species, H-*ORFs* from hermaphroditic unionoids are longer and display relatively low levels of sequence similarity to F-*ORFs* from closely related (congeneric) gonochoric species (Supplementary Figure 2). H-*ORFs* also contain repeat units not found in any of the F-*ORFs* from gonochoric species, and they sometimes possess different hydropathy profiles from the F-*ORF* proteins (e.g., *U. imbecillis* vs. *U. peninsularis*). All repetitive sequences in H-*ORFs* can be easily identified as a portion of the corresponding F-*ORF*, indicating internal duplication. No TPR or PPR motifs were found, and sequences are generally too short to correspond to these motifs. One possible mechanism that could be responsible for internal duplication of repeats independently in the various H-*ORF* sequences is DNA slippage due to the formation of DNA hairpins, a common mechanism in the creation of nucleotide repeats and thus short protein repeats [71, 72]. These independently acquired distinctive features of the five H-*ORFs* could indicate changes of function from that of the homologous F-*ORFs* in gonochoric species. Another important change in the switch from

gonochorism to hermaphroditism is the loss of the M-type mtDNA. The switch from DUI to SMI has led to the hypothesis that M- and F-type mtDNA, and particularly F-*ORF* and M-*ORF* proteins likely have coordinated roles in maintaining gonochorism [40].

Proteins that contain tandem repeats are frequently involved in interactions with other proteins or ligands such as DNA or RNA (e.g. [72, 73]). A classic example involving proteins expressed in organelles is the pentatricopeptide repeat (PPR) protein family. PPR proteins contain tandemly repeated sequences that can vary in number and are known to have roles in transcription, RNA processing, splicing, stability, editing, and translation, i.e., processes that are important for expression of organelle genomes and organelle biogenesis (see [73] for a review). In mitochondria, defects in PPR protein function can yield phenotypes associated with organelle dysfunction [73]. Interestingly, PPR proteins are key elements of the only known sex determination system in which the mitochondrial DNA is directly involved, i.e., in hermaphroditic angiosperm plants exhibiting cytoplasmic male sterility (CMS) [73]. In this nuclear-cytoplasmic sex determining system, PPR proteins appear to function as nuclear-encoded restorers of fertility, which suppress mtDNA-encoded factors that induce the inability to produce viable pollen [73]. Angiosperm plants with mutant PPR proteins are unable to produce viable pollen. Although speculative, it has been hypothesized that in freshwater mussel species with DUI, the loss of the M mitochondrial genome and macromutations in the F-*orf* gene (i.e., acquisition of tandemly repeated units in the H-*ORF* protein) could induce the ability of females to produce sperm (and eggs), leading to hermaphroditism [42].

Section 2.4.2: Conserved motifs and domains: mitochondrial export of ORFan proteins

In their previous *in silico* characterization of M-*ORFs* and F-*ORFs* of DUI bivalves belonging to the orders Mytiloida, Veneroida, and Unionoida, Milani et al. [44] used the program HHpred and found the same first four hits as presented here: LPXTG-motif cell wall anchor domain, outer membrane insertion C-terminal signal (both involved in cell membrane/surface anchoring), X-X-X-Leu-X-X-Gly heptad repeats (implicated in transcription) and PPRs (involved in post-transcriptional processes). In this unionoid-specific study we found these same four hits in all F-*ORFs*, M-*ORFs*, and H-*ORFs*, plus one hit involved in cleavage/methylation and protein transport (prepilin-type processing-associated H-

X9-DG), and one hit that serves as a recognition sequence for protein sorting and cleavage (Gly-Gly-CTERM) (Supplementary Tables V and VI). Other frequently recurring hits were found for M-*ORFs* (lysine-rich profile and nuclear localization signal), F-*ORFs* (outer membrane protein motifs), and H-*ORFs* (outer membrane or envelope proteins) (Supplementary Tables VII and VIII) As in Milani et al. [44], the regions covered by the hits are too short to perform the functions (incomplete matches), but they are supported by other hits as well (Supplementary Tables VII and VIII). Many hits to proteins with functions related to protein transport and movement, different from the ones listed above and with lower significance, are also found using profile searches for M-*ORF*s, but not for F-*ORF*s.

So far, the protein products of the F-*orf* and M-*orf* genes in unionoids have been studied only in the species *Venustaconcha ellipsiformis* [40]. Using immunoelectron microscopy, the F-*ORF* protein has been localized not only to egg mitochondria, but also on the nuclear envelope and in the nucleoplasm of unfertilized eggs [40]. Interestingly, the F-*ORF* protein has also been found on the inner mitochondrial membrane of some sperm mitochondria, which are thought to contain only M mtDNA [74]. Although the subcellular localization of the M-*ORF* protein has not been studied yet, our *in silico* detection of nuclear localization signals in several M-*ORF* sequences (Supplementary Tables VII and VIII), and of hits related to protein movement using their profile (Supplementary Tables IX and X), is consistent with the hypothesis that this protein is also exported from the organelle. Such results have already been observed in the venerid clam *Ruditapes philippinarum*, where the M-*ORF* protein has been immunolocalized both in mitochondria and in the nucleus of sperm [46]. Hence, mitochondrial ORFan proteins in DUI bivalves likely have a role (or multiple roles) in different cellular compartments ([40, 44, 46], present study), explaining the existence of functional domains that allow them to interact with several cellular elements such as membranes, cytoskeleton (see below), proteins, and nucleic acids.

The M-*ORF* and F-*ORF* of *V. ellipsiformis* are 25-30 kDa each [40], and thus may be able to diffuse into the nucleus through pores in the nuclear envelope, which are relatively large and flexible [75], but their export from the mitochondria remains unexplained. Mitochondrial import of proteins is a well-known process [76], but the opposite process of mitochondrial export is still largely obscure (e.g., [77]). The export of cell death effectors [78],

retrograde signals *humanin* and MOTS-c [79], and small peptides to trigger retrograde nuclear signalling in mitochondrial unfolded protein response are all somewhat characterized, but mitochondrial protein export is still largely unstudied, particularly for larger molecules [80]. One speculative hypothesis would be a system akin to autotransporters, which use the Type V secretion system in Gram-negative bacteria, including the proteobacteria (the presumed ancestor of mitochondria) [81]. Classical autotransporters, which are important virulence factors in many Gram-negative pathogens, are known to pass through the inner membrane of Gram-negative bacteria with the help of their N-terminal signal peptide, whereas their C-terminal motif forms a beta-barrel pore in the outer membrane to allow the secretion of the passenger domain (the mature protein) [81]. Interestingly, it has been demonstrated that the evolutionary conservation in the biogenesis of beta-barrel proteins allows mitochondria to assemble bacterial autotransporters in their functional form [82]. However, autotransporters are usually extremely large molecules, and although F-*ORFs* and H-*ORFs* contain a similar 3-domain structure with an N-terminal SP, a C-terminal "motif", and a central variable region, they are much shorter and do not possess beta-barrel structures. They cannot, therefore, behave exactly like autotransporters, however we can hypothesize that the ORFan proteins could use their N-terminal SP to pass through the inner membrane, and their C-terminal motif to traverse the outer membrane. Their export could potentially be facilitated by proteins belonging to the evolutionarily conserved Omp85 family, which are essential for outer membrane biogenesis in mitochondria and chloroplasts, and recognize species-specific C-terminal motifs when functioning as assembly factors and in protein export in bacteria [83, 84]. Clearly, further studies are needed to better understand the process of mitochondrial export not only in bivalves, but in other animal species as well.

Section 2.4.3: Putative origin for freshwater mussel mitochondrial ORFans

As mentioned above, previous *in silico* analyses provided many clues consistent with a viral origin of DUI bivalve mitochondrial ORFans, even if the probability of the hits were sometimes low and the regions of similarity of short length [44]. Except for two sequences (M-*ORF* of *Cumberlandia monodonta* and F-*ORF* of *Anodonta anatina*), our results revealed the presence of at least one viral hit for each sequence analyzed, also with low probability values and short regions of similarity. Although not significant, the same viral hit

(neuraminidase) has been found multiple times for the F-*ORF* profiles, and a few hits related to viral activities, or insertion of foreign elements, were also retrieved for M-*ORF* profiles. As in Milani et al. [44], many hits pointing to immune system, defense, and antigens were found (Figure 4, Table VI and Supplementary Tables XI-XLV), supporting the hypothesis that mitochondrial ORFans in bivalves may have originated from viral elements with a function in immune response and apoptosis control. However, H-*ORFs* returned more hits to viral proteins than M-*ORFs* or F-*ORFs* which, given that they are likely derived from F-*ORFs*, suggests that these genes may be converging towards molecular properties similar to viral proteins. In addition, for each sequence analyzed, we also obtained hits with stronger probability values for bacterial or metazoan genes, and with functions other than immune response (Table VI and Supplementary Tables XI-XLV). Our objective was to compare more closely related sequences to obtain clearer patterns that could help to better understand the origin(s) and function(s) of these mitochondrial ORFans in bivalves. However, these ORFans evolve at a rate that limits our ability to fully characterize their function and/or evolutionary origins. In general, rapidly evolving ORFan genes for which homologues cannot be determined easily have been implicated in lineage-specific processes, such as immune system functioning and sex determination, and have been proposed to be major contributors to the origin of adaptive evolutionary novelties [85, 86]. It is noteworthy that collectively, studies of new genes in animal species have ascribed the testis as having a central role in the process of gene birth and evolution [86].

The birth of new genes involves a variety of mechanisms, including the origin of new protein-coding and RNA genes from previously nonfunctional genomic sequences, various types of gene fusions, and the formation of new genes from RNA intermediates [86, 87]. However, gene duplication is thought to be the mechanism underlying the origin of most novel genes, and thus represents one of the most important processes for functional innovation during evolution [86]. Interestingly, several of our sequences analysed returned hits to proteins involved in mitochondrial energy production, including proteins of the electron transport system, suggesting that duplication and neofunctionalization of a mitochondrial gene could be the source of freshwater mussel mitochondrial ORFans. Several M-*ORF* sequences (i.e., 6 species out of 9: *C. monodonta, I. japanensis*, *H. menziesii*, *Q. quadrula, S. carinatus*, *V.*

ellipsiformis) returned hits, some of them with high probability values, to the subunit *ATP8* of the mitochondrial ATP synthase complex V (Table VII), and M-*ORF* profiles to subunit b of bacterial ATP synthase. These results are interesting in at least two ways. Firstly, because the *atp8* and M-*orf* genes are localized one beside another in the M mitochondrial genome, in a region corresponding to one of the three gene order rearrangements observed between female and male mtDNAs in freshwater mussels with DUI [40], and secondly, because the *atp8* gene is highly modified or reported missing in other bivalve species with DUI due to its short length and rapid evolution causing difficulties in annotation (e.g., [88–90]), leading to the possibility that this gene acquired DUI-specific functions that hamper its annotation in DUI bivalves. It is highly conceivable that a duplication event, as described in several other animal mtDNAs [91], of the region containing the *atp8* gene happened in an ancestral freshwater mussel species with DUI, allowing one of the two duplicate *atp8* copies to evolve new male-specific functional properties and thus giving birth to the M-*orf* gene in M-type mtDNAs. Considering this, both mitochondrial *ATP8* and bacterial subunit b hits may indicate a mitochondrial localization for M-*ORF* in the F0 subunit of complex V, the part of ATP synthase where protons pass through the inner membrane from the intermembrane space to the matrix. Examples of mtDNA-encoded non-canonical subunits of the F0 complex are already known from studies on protists [92], and unionoid M-*ORF*s might be a metazoan variant of this scenario. However, how and if the M-*ORF* in these species could alter the membrane potential by locating itself in complex V, and possibly drive sperm mitochondria inheritance by such mechanism (as proposed by [93]) will be questions for future studies.

Individual F-*ORF* sequences also returned many hits pointing to mitochondrial membrane proteins, often *NAD2*, although with relatively high E-values. Nonetheless, this is an interesting result because the *nad2* and F-*orf* genes are typically localized beside one another in the female mitochondrial genome, in a region corresponding to the only gene order rearrangement observed among F mtDNAs in freshwater mussels with DUI [42]. It is plausible that this region could have been subjected to a duplication event and subsequent adaptation of one of the two copies of *nad2*. It is worth noting that the *nad2* gene is also localized beside the F-*orf* gene in the marine clam *Ruditapes philippinarum* [90] (this, however, is not the case for other species with DUI). Finally, and not surprisingly, all H-*ORF*

sequences returned hits to F-*ORF* sequences (Table 7), and many hits for F-*ORF* profiles are annotated H-*ORF*s (Supplementary Tables IX and X), supporting previous results that the former gene derived from the latter [40].

With such a rapid rate of evolution, it would not be unreasonable for the mitochondrial ORFans to rapidly lose their resemblance to the highly conserved mitochondrial genes from which they evolved. Our results do not refute the hypothesis that these ORFans may originate from viral sequences, but they open up the possibility of a mitochondrial origin. Additionally, the differences in the genomes across the orders Mytiloida, Veneroida, and Unionoida, when coupled with the fact that *ATP8* and *NAD2* are the most likely candidates for duplications leading to the development of M-*ORF* and F-*ORF,* respectively suggest that this mechanism would best fit multiple origins of DUI and the necessary factors. On the other hand, the rate of evolution of these genes and the structural and functional similarities seen across orders hints at a single origin with significant modifications over the course of the evolution of these orders. Sex determination factors are wildly diverse and tend to evolve rapidly despite strong conservation of the molecular pathways that they trigger [94], thus it is likely that extensive further work will be needed before a strong theory on the origin(s) of DUI can be built.

Section 2.4.4: Predicted functions for freshwater mussel mitochondrial ORFans

Overall, when we consider the most frequently recurring categories of hits, the functions of proteins bearing sequence and/or structural similarities to the freshwater mitochondrial ORFans follow a general pattern. For M-*ORFs*, these include membrane associated proteins, transport, and cellular processes, i.e., signalling, cell cycle control, and cytoskeleton organization (cell differentiation during development). For F-*ORFs*, they mainly include membrane-associated proteins, trafficking and transport, and the immune system. For H-*ORF*s*,* hits primarily returned structural proteins, especially glycoproteins such as collagen and collagen-like proteins, membrane-associated proteins, signalling, transport, and transcription. All mitochondrial ORFans in freshwater mussels seem capable of DNA, RNA, and protein binding, and transcription regulation.

Milani et al. [44] suggested that the novel ORFs might have a role in producing the aggregated and dispersed patterns of distribution of spermatozoon mitochondria observed in

early male and female embryos, respectively. This hypothesis has been supported by a subsequent study of the M-*ORF* in the marine clam *Ruditapes philippinarum* (named RPHM21 in this species) [46], in which the prediction of domains involved in cytoskeleton interactions, as well as the localization of the RPHM21 product in sperm mitochondria and around the animal-vegetal axis of embryos, supported a role of this protein in the distribution pattern of spermatozoon mitochondria observed in DUI embryos. The results obtained in the present study for the freshwater mussel M-*ORFs* also provide support for this hypothesis. Although it remains unclear if mitochondrial ORFans in freshwater mussels originated from viral or mitochondrial elements, our results for M-*ORFs* reveal connections with the cytoskeleton, such as microtubule-binding proteins, actin-binding proteins and proteins with a role in protein-cytoskeleton interactions (e.g., ankyrin). With their predicted SPs and TM helices, M-*ORFs* may be targeted to sites outside sperm mitochondria and be responsible for their cellular distribution and positioning in developing embryos. It has been suggested that mitochondrial dynamics, including motility, fusion, fission, and autophagy, must be, at least partly, controlled by "signalling" from the respective individual mitochondrion [95]. Although no protein of the dynamics machinery has been identified in bivalves yet, the mitochondrially-encoded M-*ORF* in bivalves with DUI represents an ideal candidate for direct control of sperm mitochondria.

A possible viral origin of the M-*orf* gene, as previously suggested [44, 46], supports the hypothesis of a role for its protein product during embryo development, that is to prevent the recognition of sperm mitochondria by the degradation machinery in DUI zygotes, as some viral proteins do in the immune recognition pathway, thus explaining the acquired capability of sperm mitochondria to avoid degradation and invade the germ line. Milani et al. [46] also described several retroviral genes co-opted by the host which have been shown to have roles in early development and in sex-specific functions, such as gamete differentiation and reproduction. This supports the hypothesized connection between DUI, the novel ORFs, and the maintenance of separate sexes in freshwater mussels [40]. Our results do not refute the hypothesis that the mitochondrial ORFans in bivalves with DUI might have a viral origin, but irrespective of this putative viral endogenization, a reproductive role for ankyrin-like proteins, to which freshwater mussel M-*ORFs* showed structural similarities (see Supplementary Tables

XI-XIX), is already well established (e.g. [96]). Remarkably indeed, Yu et al. [96] reported that a mitochondrially-localized ankyrin repeat protein (ANK6) is essential for fertilization in *Arabidopsis*, specifically for gamete recognition, possibly by regulating mitochondrial gene expression.

A possible function in reproduction, fertilization, gamete development or sex determination (and regulation of mitochondrial expression) is an attractive hypothesis for the mitochondrial ORFans in freshwater mussels. For example, the F-*ORF* protein could participate in the inhibition of testicular development in embryos that will become females, and the extreme modifications seen in H-*ORFs* (highly modified versions of F-*ORF* proteins) could explain why development of some testicular tissue is not inhibited in hermaphrodites. Interestingly, the F-*ORF* protein has not only been localized in the mitochondria, nuclear membrane, and nucleoplasm of eggs, but also in some sperm mitochondria [40, 74]. Because small proteins like this often diffuse into the nucleus without a specific targeting signal, the nuclear localization may not be specific, however, mitochondrial localization depends on a signal peptide at the N terminus of the protein [96]. Because the F mtDNA is not present in DUI bivalve sperm mitochondria [97], either there is a version of the F-*orf* gene in the nuclear genome, or the F mtDNA-encoded F-*ORF* protein is exported from F-type mitochondria and imported, with the help of its N-terminal signal peptide, into M-type mitochondria, where it could regulate mitochondrial gene expression, for example of the M-*orf* gene. Future experiments and examination of a freshwater mussel nuclear genome, which is currently underway in our laboratory, are needed to verify these hypotheses. But as hypothesized by Milani et al. [44], the M-*ORF* protein could be a masculinizing factor and that sperm from different males could carry different amounts of transcript and/or protein, determining the quantity of protein in embryos thus shifting development toward maleness. Although speculative, the F-*ORF* protein in sperm, possibly with the help of some nuclear-encoded factors, could be responsible for the regulation of this process. Yusa et al. [98], in their DUI sex-determination model, predicted the existence of such secondary or minor sex-determining mitochondrial factors.

Although the expression of the H-*ORF* protein in hermaphrodites has not yet been experimentally confirmed, the high level of amino acid sequence and structural similarities

within species suggest that it is functional. H-*ORFs* are predicted to be membrane-associated and/or secreted, and may act in signalling and transcription. Our results suggest that these proteins may in fact be glycoproteins given the large number of high probability and high coverage hits to glycoproteins (see Supplementary Tables XXXII-XLV). Interestingly, a previous study of the reproductive tracts of *Utterbackia imbecillis* identified an unknown carbohydrate or glycoprotein co-localized with cells determined to be secretory, which was absent from closely-related gonochoric species [99]. To date, this molecule has not been characterized, but the authors suggested that it might inhibit self-fertilization [99]. This molecule should be further investigated to determine whether it could be the H-*ORF* protein and determine if it has a role in fertilization.

Section 2.4.5: Conclusions and future directions

Knowledge of metazoan mitochondrial genomes is constantly changing, expanding and moving ever further from the norms of the past (e.g. [4]). One major deviation from mitochondrial paradigms is the DUI system with its novel, lineage-specific mitochondrial ORFans. Considering previous data [44, 46] and the data presented here, it is clear that the striking similarities observed among mitochondrial ORFans in distantly-related bivalve species indicate some commonality, that is a role for these genes in the DUI mechanism. Our results also lead to some clear questions for future work: are mitochondrial ORFans in freshwater mussels (and other DUI bivalves) of viral or mitochondrial origin? Are H-*ORFs* expressed in hermaphroditic species? Are they glycoproteins - or more specifically - the molecule observed by Henley et al. [99]? What are the subcellular localizations of M-*ORFs* and H-*ORFs*? Are the proteins exported, and if so, do the SP and C-terminal motif play essential roles in this? As recently proposed [100], DUI is an intriguing system to look for antagonistic interactions between distorting mitochondria and nuclear suppressors similar to CMS in plants. If the F-*ORF* and M-*ORF* proteins are indeed antagonistic molecules, i.e., with the F-*ORF* participating in the inhibition of testicular development in female developing embryos and the M-*ORF* participating in the inhibition of ovarian development in male developing embryos, this could explain why macromutations in the F-*ORF* protein (that turns it into a H-*ORF*) would allow for testis development in otherwise female gonads (i.e.,

hermaphroditism). Most importantly, the mechanisms underlying DUI and sex determination in bivalves remain to be elucidated.

Chapitre 3 : Discussion, perspectives et conclusion

Les connaissances concernant l'ADN mitochondrial chez les espèces animales changent rapidement et plusieurs études ont permis la découverte d'une multitude d'exceptions aux normes établies par le passé (voir [4] pour une revue récente). Parmi ces exceptions se trouve le système de la DUI et ses ORFans mitochondriaux spécifiques aux sexes chez les bivalves. Les données présentées ici et dans les études antérieures [44, 46] indiquent des similarités remarquables parmi les ORFans des différents ordres de bivalves, et suggèrent un rôle pour ces gènes dans la DUI.

La première étude *in silico* des ORFans mitochondriaux chez les espèces à DUI incluait des espèces très distantes, et les résultats ont montré qu'il y a plus de similarité structurale et fonctionnelle à l'intérieur d'un ordre qu'entre les trois ordres étudiés (Mytiloida, Veneroida, Unionoida) [44]. Ces résultats, qui laissent ouverte la question d'une origine unique ou multiple pour la DUI, ont poussé les auteurs à proposer des études sur des espèces plus proches phylogénétiquement afin d'avoir une meilleure idée de l'origine et des fonctions des nouveaux gènes mitochondriaux chez les espèces à DUI [44]. Pour ces raisons, et aussi parce que l'existence d'espèces hermaphrodites avec des F-*ORF* hautement modifiés amène un autre aspect intéressant, nous avons décidé de faire des analyses *in silico* avec les séquences d'ORFans mitochondriaux d'espèces de moules d'eau douce (ordre Unionoida) pour mieux comprendre leurs origines et fonctions.

Section 3.1 : Conservation des séquences et structures des ORFans

Nos analyses démontrent que les séquences des M-*ORF*s sont très variables au niveau inter-espèce : leur longueur varie beaucoup, et les séquences en acides aminés sont peu similaires. Malgré cette grande variabilité, leurs structures secondaires semblent être bien conservées. Quant aux protéines F-*ORF*s, leurs séquences sont beaucoup plus conservées, surtout au C-terminus. Elles ont toutes une TMH et un SP prédits dans leur portion N-terminale. Donc, malgré les divergences importantes observées au niveau des séquences en acides aminés entre les espèces, les protéines M-*ORF*s et F-*ORF*s semblent être conservées au niveau de la structure, ce qui suggère une conservation inter-espèces de leurs fonctions aussi.

Contrairement aux M-*ORF*s et F-*ORF*s, les protéines H-*ORF*s sont plus longues et contiennent beaucoup de répétitions en tandem. La séquence protéique H-*ORF* de chaque espèce hermaphrodite est relativement similaire aux F-*ORF*s des espèces proches-parentes, la divergence étant surtout due aux séquences répétitrices. Tout comme les autres ORFans, les protéines H-*ORF*s ont une TMH prédite au N-terminus et un SP aussi. Un mécanisme possiblement responsable pour les répétitions présentes dans les H-*ORFs* est le glissement de l'ADN dû à la formation de structures secondaires de type « épingles à cheveux » durant la réplication, ce qui est fréquemment à l'origine des séquences répétitrices [71, 72]. Les protéines H-*ORF*s des 5 espèces hermaphrodites étudiées montrent des changements structuraux similaires, ce qui pourrait indiquer un changement de fonction par rapport aux protéines F-*ORF*s. Il a déjà été proposé que parce que la perte du génome paternel M est toujours associée aux modifications des protéines F-*ORFs* (apparition des H-*ORFs*) chez les hermaphrodites, il est fort probable que le rôle des protéines F-*ORF* et M-*ORF* est de maintenir le gonochorisme (les sexes séparés) et que leur perte/modification coordonnée mène à la production d'individus hermaphrodites [40].

Section 3.2 : Exportation des ORFans de la mitochondrie

Jusqu'à présent, le seul unionidé dont les ORFans mitochondriaux avaient été le sujet d'étude plus poussée est *Venustaconcha ellipsiformis* [40]. À l'aide de techniques immunohistochimiques et de la microscopie électronique, la protéine F-*ORF* a été détectée non seulement dans les mitochondries de type F, mais également dans l'enveloppe nucléaire et le nucléoplasme des œufs non-fertilisés [40]. De plus, cette même protéine a été localisée sur la membrane interne de certaines mitochondries de spermatozoïdes, qui devraient en temps normal contenir exclusivement des mitochondries de type M [74]. La localisation subcellulaire de la protéine M-*ORF* n'est pas connue, mais les analyses présentées ici ont détecté des signaux de localisation nucléaire dans plusieurs séquences M-*ORF*, ce qui suggère que la protéine pourrait être exportée. Des telles résultats ont été trouvés dans un étude de la palourde *Ruditapes philippinarum*, où M-*ORF* (RPHM21) a été détectée dans les mitochondries et les noyaux des spermatozoïdes [46]. Donc, les ORFans ont probablement un rôle (ou plusieurs rôles) dans divers compartiments cellulaires ([40, 44, 46], cette étude), ce qui explique les

prédictions des domaines fonctionnels qui permettent l'interaction avec divers éléments de la cellule, tels que les membranes, le cytosquelette, les protéines et les acides nucléiques.

Présentement, il y a peu d'hypothèses concernant l'exportation de grandes protéines de la mitochondrie. Un de ces hypothèses implique des vésicules dérivées de la mitochondrie formées lors d'un stress oxydatif, tandis qu'une autre propose des translocases mitochondriales, telles l'insertase *Oxa-1*, qui sont normalement utilisées pour importer des protéines, mais qui pourraient fonctionner en sens inverse aussi [77]. Aucun de ces mécanismes n'a été observé jusqu'à présent. Une autre possibilité, quoique très spéculative, est que des protéines exportées de la mitochondrie pourraient utiliser un mécanisme similaire aux autotransporteurs du système de sécrétion Type V chez les bactéries Gram-négatives, qui incluent les protéobactéries, ancêtres présumées des mitochondries [81]. Ces protéines autotransporteurs sont des facteurs de virulence importants chez plusieurs pathogènes, et ils passent à travers la membrane interne à l'aide d'un peptide signal N-terminal tandis qu'un long motif C-terminal forme un tonneau bêta qui permet le passage du domaine passager, qui est la protéine mature, par la membrane externe [81]. Il est intéressant de noter que les mitochondries peuvent assembler des autotransporteurs dans leur forme fonctionnelle [82]. Cependant, ces protéines sont généralement très grandes avec une architecture répétitive, et bien qu'une bonne partie des séquences étudiées ici ont une structure à 3 domaines avec un SP N-terminal, un motif C-terminal et une région centrale variable, elles sont beaucoup trop courtes et ne contiennent pas le tonneau bêta qui pourrait leur permettre de fonctionner comme les autotransporteurs classiques. Par contre, ces protéines ORFans pourraient utiliser leurs motifs N-terminal et C-terminal pour passer à travers les deux membranes mitochondriales à l'aide des protéines de la famille Omp85, par exemple, qui sont essentielles pour la biogénèse de la membrane externe des mitochondries et des chloroplastes, et qui reconnaissent des motifs C-terminaux spécifiques aux espèces lorsqu'impliquées dans l'exportation de protéines chez les bactéries [83, 84]. Il serait donc intéressant d'introduire ce gène dans des bactéries, le faire exprimer et vérifier si les bactéries secrètent la protéine et, le cas échéant, comment (ce qui pourrait être fait, par exemple, avec des souches ayant certains composants des systèmes de sécrétion silencés). Une telle étude pourrait également éclaircir les rôles des SP N-terminal et motif C-terminal retrouvés dans les protéines F-*ORF*.

Section 3.3 : Fonctions potentielles

Nous proposons l'hypothèse que la protéine F-*ORF* pourrait participer à l'inhibition du développement des cellules spermatogéniques chez les embryons qui deviendront des femelles, et que les importantes modifications présentes dans les H-*ORFs* chez les hermaphrodites pourraient expliquer leur capacité de produire des spermatozoïdes. Tel que mentionné ci-haut, la protéine F-*ORF* a été trouvée non seulement dans les mitochondries, la membrane nucléaire et le nucléoplasme des œufs, mais aussi dans certaines mitochondries des spermatozoïdes chez l'espèce *Venustaconcha ellipsiformis* [40, 74]. Puisque des petites protéines peuvent arriver au noyau par diffusion sans nécessiter de signal particulier, la localisation nucléaire n'est pas nécessairement spécifique, cependant la localisation mitochondriale dépend d'un SP [96]. Étant donné que l'ADNmt de type F n'est pas présent dans les mitochondries des spermatozoïdes chez les moules d'eau douce [97], soit il y a une copie nucléaire du gène F-*orf*, soit la protéine est exportée des mitochondries de type F et importée dans les mitochondries de type M. La protéine F-*ORF* pourrait par exemple réguler l'expression des gènes mitochondriaux paternels tels que le M-*orf*. Des expériences biochimiques ainsi que la séquence d'un génome nucléaire complet seront nécessaires pour vérifier ces hypothèses.

La protéine M-*ORF* quant à elle, pourrait être un facteur de masculinisation et, tel que proposé par Milani et al. [44], les spermatozoïdes chez les bivalves à DUI pourraient contenir différents niveaux du produit du gène, ce qui déterminerait le sexe de l'embryon. La protéine F-*ORF* pourrait être responsable de réguler ce processus dans les spermatozoïdes. L'existence des facteurs mitochondriaux secondaires dans la détermination du sexe a déjà été prédite pour les bivalves à DUI [98]. On ne sait pas encore si le gène H-*orf* est exprimé, mais la conservation intra-espèce des séquences en acides aminées et les similarités structurales inter-espèces suggèrent qu'il est fonctionnel. Même si les protéines H-*ORF*s sont très divergentes des protéines F-*ORF*s des espèces proches parentes – surtout à cause des séquences répétitrices – les prédictions fonctionnelles sont similaires à 60% au niveau des catégories générales. Les prédictions pour les protéines H-*ORF*s incluent des protéines associées à la membrane et/ou sécrétées, et ayant des rôles dans la signalisation et la transcription. Nos résultats indiquent que ces protéines pourraient, en fait, être des glycoprotéines, et une étude

antérieure du système reproducteur de l'espèce hermaphrodite *Utterbackia imbecillis* a identifié un carbohydrate ou une glycoprotéine dans ou autour de cellules sécrétrices, mais absent chez une espèce proche-parente à sexes séparés [99]. Cette molécule n'a jamais été caractérisée, mais les auteurs ont suggéré qu'elle pourrait prévenir l'autofécondation. Il serait intéressant de vérifier si cette molécule pourrait être la protéine H-*ORF* codée par le génome mitochondrial.

Section 3.4 : DUI et conflit génomique

Le génome mitochondrial de type M a déjà été décrit comme « presque égoïste » dans le sens qu'il doit remplir les fonctions normales de l'ADNmt dans les cellules spermatogéniques, mais pas dans les autres tissus [21]. Cette hypothèse peut expliquer l'évolution rapide de ce génome, et particulièrement le gène M-*orf*, qui n'aurait peut-être aucun rôle dans la production d'énergie mais seulement dans la détermination du sexe. En bref, le gène M-*of* pourrait être un composant égoïste d'un génome mt « presque égoïste ». Les nombreuses divisions cellulaires durant la gamétogénèse ainsi que les dommages oxydatifs subits dans les mitochondries des spermatozoïdes sont également des sources de mutations possibles [21]. Une relaxation de la sélection purificatrice qui est normalement très forte pour les génomes mitochondriaux a aussi été proposée pour expliquer l'évolution rapide du génome M [101]. Les taux de substitutions sont particulièrement élevés aux sites non-synonymes (voir références dans [21]). Il a déjà été proposé que ces changements puissent avoir des avantages au niveau de la performance des spermatozoïdes, mais les résultats expérimentaux ne sont pas entièrement conclusifs [102, 103].

Le gène F-*orf* pourrait également présenter une source additionnelle de pression sélective. En effet, les similarités entre les résultats obtenus pour les protéines F-*ORFs*, M-*ORFs* et H-*ORFs* sont frappantes et suggèrent que ces protéines ont des fonctions similaires chez les espèces de moules d'eau douce. Chez les espèces à sexes séparés, il semble que ces gènes aient des fonctions similaires dans les mêmes parties de la cellule, et donc, ils pourraient potentiellement être en conflit. Par exemple, des molécules antagonistes devraient évoluer plus rapidement pour gagner un avantage, ou simplement survivre (tout comme les idées de la course aux armements évolutionnaires et l'hypothèse de la reine rouge, e.g., [104]). Le besoin

d'évoluer rapidement dans un cas de conflit pourrait également expliquer le taux de mutations non-synonymes plus élevé – pour survivre dans une telle situation, il faut subir des changements qui dépassent le niveau de la séquence nucléotidique.

Section 3.5 : Unionoida

Quant au Unionoida, comment et pourquoi les hermaphrodites évoluent restent à être élucidé. L'étude d'une population en transition vers l'hermaphrodisme pourrait éclarcir la situation en identifiant des états intermédiaires de plusieurs caractéristiques importantes, incluant la transition du gène F-*orf* vers H-*orf* et l'expression de la protéine H-*ORF*. L'obtention des génomes mitochondriaux d'hermaphrodites accidentels serait aussi un atout pour une meilleure compréhension de la fonction des nouveaux gènes. Une plus grande banque des séquences (incluant des gènes nucléaires) sera importante pour mieux comprendre le lien entre le système DUI et la détermination du sexe chez les moules d'eau douce. Finalement, le développement de techniques pour croiser des unionidés en laboratoire permettrait des tests de compatibilité entre espèces proche-parentes : par exemple, on pourrait féconder l'œuf d'une femelle avec le sperme d'un hermaphrodite, et l'œuf d'un hermaphrodite avec le sperme d'un mâle pour voir s'ils sont compatibles et, le cas échéant si la DUI et les sexes séparés persistent chez les descendants.

Nos résultats amènent également beaucoup de nouvelles questions: d'abord, quelle est la véritable origine des ORFans chez les espèces à DUI? Est-ce que le gène H-*orf* s'exprime? Des résultats préliminaires d'une étude en cours dans notre laboratoire ont montré la présence de la séquence H-*orf* dans le transcriptome de *U. imbecillis*, suggérant que oui. Nos résultats suggèrent également que cette protéine serait une glycoprotéine (probablement structurale), et donc la purification et caractérisation de celle-ci pour vérifier ceci, ainsi que l'étude de sa localisation subcellulaire pourrait aider à déterminer si la glycoprotéine identifiée par Henley et al. [99] est belle et bien la protéine H-*ORF*. En fait, une étude plus approfondie de la localisation subcellulaire des H-*ORF*s, M-*ORF*s et F-*ORFs* est nécessaire chez les unionidés.

Les différences parfois mineures entre les F-*ORF*s et H-*ORF*s mènent à quelques questions aussi. Tout d'abord, qu'est-ce qui cause la transition vers l'hermaphrodisme? Des facteurs environnementaux ont déjà été proposés, mais généralement peu étudiés en

profondeur [38, 105–110]. Le cas des espèces du genre *Margaritifera* est particulièrement intéressant, parce que la protéine H-*ORF* de *M. falcata* diffère de la protéine F-*ORF* de *M. margaritifera* par une seule répétition, ce qui met en question à quel point les séquences doivent changer pour produire un hermaphrodite. Vu qu'une seule répétition semble suffisante, il serait intéressant d'investiguer si cette répétition seule peut déclencher des changements, ou si elle est, en fait, une conséquence de l'hermaphrodisme, ou bien si des facteurs nucléaires sont également nécessaires. Le séquençage de génomes nucléaires complets et l'étude des différences entre les transcriptomes des hermaphrodites et des espèces à sexes séparés qui sont présentement en cours dans notre laboratoire nous permettront de mieux caractériser le système de la DUI et son lien potentiel avec la détermination du sexe.

Finalement, avec trois génomes, et deux gènes mitochondriaux spécifiques au sexe ayant des fonctions similaires, la DUI pourrait représenter le premier cas de détermination du sexe impliquant directement les mitochondries chez les animaux. Ce système de transmission mitochondriale unique s'avèrera certainement être un cas riche en découvertes sur les conflits intergénomiques, d'un point de vue mitonucléaire, et également entre les sexes.

Bibliographie

1. McBride HM, Neuspiel M, Wasiak S: **Mitochondria: more than just a powerhouse.** *Curr Biol* 2006, **16**:R551–60.

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

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

4. Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M: **A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs**. *Trends Genet* 2014, **30**:555–564.

5. Kayal E, Bentlage B, Collins AG, Kayal M, Pirro S, Lavrov D V.: **Evolution of linear mitochondrial genomes in medusozoan cnidarians**. *Genome Biol Evol* 2012, **4**:1–12.

6. Doublet V, Souty-Grosset C, Bouchon D, Cordaux R, Marcadé I: **A thirty million year-old inherited heteroplasmy.** *PLoS One* 2008, **3**:e2938.

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

8. Wu X, Li X, Li L, Xu X, Xia J, Yu Z: **New features of Asian Crassostrea oyster mitochondrial genomes: a novel alloacceptor tRNA gene recruitment and two novel ORFs**. *Gene* 2012, **507**:112–8.

9. Milbury CA, Gaffney PM: **Complete mitochondrial DNA sequence of the eastern oyster Crassostrea virginica**. *Mar Biotechnol* 2005, **7**:697–712.

10. Shao Z, Graf S, Chaga OY, Lavrov D V.: **Mitochondrial genome of the moon jelly Aurelia aurita (Cnidaria, Scyphozoa): A linear DNA molecule encoding a putative DNA-dependent DNA polymerase**. *Gene* 2006, **381**:92–101.

11. Fukami H, Knowlton N: **Analysis of complete mitochondrial DNA sequences of three members of the Montastraea annularis coral species complex (Cnidaria, Anthozoa, Scleractinia)**. *Coral Reefs* 2005, **24**:410–417.

12. Flot JF, Tillier S: **The mitochondrial genome of Pocillopora (Cnidaria: Scleractinia) contains two variable regions: The putative D-loop and a novel ORF of unknown function**. *Gene* 2007, **401**:80–87.

13. Haen KM, Lang BF, Pomponi SA, Lavrov D V.: **Glass Sponges and Bilaterian Animals Share Derived Mitochondrial Genomic Features: A Common Ancestry or Parallel Evolution?** *Mol Biol Evol* 2007, **24**:1518–1527.

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

15. Hoeh WR, Blakley KH, Brown WM: **Heteroplasmy suggests limited biparental inheritance of Mytilus mitochondrial DNA.** *Science (80-)* 1991, **251**:1488–90.

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

17. Zouros E, Freeman KR, Ball AO, Pogson GH: **Direct evidence for extensve paternal mitochondrial DNA inheritance in the marine mussel _Mytilus_** . *Nat* 1992, **359** :412–414.

18. Skibinski DO, Gallagher C, Beynon CM: **Sex-limited mitochondrial DNA transmission in the marine mussel Mytilus edulis.** *Genetics* 1994, **138**:801–809.

19. Obata M, Sano N, Kawamura K, Komaru A: **Inheritance of two M type mitochondrial DNA from sperm and unfertilized eggs to offspring in Mytilus galloprovincialis**. *Dev Growth Differ* 2007, **49**:335–344.

20. Chakrabarti R, Walker JM, Chapman EG, Shepardson SP, Trdan RJ, Curole JP, Watters GT, Stewart DT, Vijayaraghavan S, Hoeh WR: **Reproductive function for a C-terminus extended, male-transmitted cytochrome c oxidase subunit II protein expressed in both spermatozoa and eggs.** *FEBS Lett* 2007, **581**:5213–9.

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

22. Dalziel AC, Stewart D.T.: **Tissue-specific expression of male-transmitted mitochondrial DNA and its implications for rates of molecular evolution in \emph{Mytilus} mussels (Bivalvia: Mytilidae)**. *Genome* 2002, **45**:348–355.

23. GarridoRamos MA, Stewart DT, Sutherland BW, Zouros E: **The distribution of male-transmitted and female-transmitted mitochondrial DNA types in somatic tissues of blue mussels: Implications for the operation of doubly uniparental inheritance of mitochondrial DNA**. *Genome* 1998, **41**:818–824.

24. Cao L, Kenchington E, Zouros E: **Differential Segregation Patterns of Sperm Mitochondria in Embryos of the Blue Mussel (Mytilus edulis)**. *Genet Soc Am* 2004, **894**:883–894.

25. Milani L, Ghiselli F, Passamonti M: **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* 2012, **318**:182–189.

26. Breton S, Beaupré HD, Stewart DT, Hoeh WR, Blier PU: **The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough?** *Trends Genet* 2007, **23**:465–474.

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

28. Hurst LD, Hoekstra RR: **Shellfish genes kept in line**. *Nature* 1994, **368**:811–812.

29. Zeh J a., Zeh DW: **Maternal inheritance, sexual conflict and the maladapted male**. *Trends Genet* 2005, **21**:281–286.

30. Zouros E: **The exceptional mitochondrial DNA system of the mussel family Mytilidae.** *Genes Genet Syst* 2000, **75**:313–318.

31. Kenchington E, Macdonald B, Cao L, Tsagkarakis D, Zouros E: **Genetics of Mother-Dependent Sex Ratio in Blue Mussels (Mytilus spp .) and Implications for Doubly Uniparental Inheritance of Mitochondrial DNA**. *Genetics* 2002, **1588**:1579–1588.

32. Hoeh WR, Frazer KS, Naranjo-Garcia E, Trdan RJ: **A phylogenetic perspective on the evolution of simultaneous hermaphroditism in a freshwater mussel Clade (Bivalvia: Unionidae: Utterbackia)**. *Malacol Rev* 1995, **28**:25–42.

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

34. Bauer G, Wächtler K (Eds): *Ecology and Evolution of the Freshwater Mussels Unionoida*. 1st edition. New York: Springer-Verlag Berlin Heidelberg; 2001.

35. Kat PW: **Sexual selection and simultaneous hermaphroditism among the Unionidae (Bivalvia: Mollusca)**. *J Zool* 1983, **201**:395–416.

36. Walker JM, Curole JP, Wade DE, Chapman EG, Bogan AE, Watters GT, Hoeh WR, Al WET: **Taxonomic Distribution and Phylogenetic Utility of Gender- Associated Mitochondrial Genomes in the Unionoida (Bivalvia)**. *Malacologia* 2006, **48**:265–282.

37. Van der Schalie H: **Hermaphroditism Among North American Freshwater Mussels**. *Malacologia* 1970, **10**:93–112.

38. Bauer G: **Reproductive Strategy of the Freshwater Pearl Mussel Margaritifera margaritifera**. *J Anim Ecol* 1987, **56**:691–704.

39. Heller J: **Hermaphroditism in molluscs**. *Biol J Linn Soc* 1993, **48**:19–42.

40. Breton S, Stewart DT, Shepardson S, Trdan RJ, Bogan AE, Chapman EG, Ruminas AJ, Piontkivska H, Hoeh WR: **Novel protein genes in animal mtDNA: A new sex determination system in freshwater mussels (Bivalvia: Unionoida)?** *Mol Biol Evol* 2011, **28**:1645–1659.

41. Stewart DT, Hoeh WR, Bauer G, Breton S: **Mitochondrial Genes , Sex Determination and Hermaphroditism in Freshwater Mussels (Bivalvia : Unionoida)**. In *Evolutionary Biology: Exobiology and Evolutionary Mechanisms*. Edited by Pontarotti P. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg; 2013:245–255.

42. Breton S, Beaupre HD, Stewart DT, Piontkivska H, Karmakar M, Bogan AE, Blier PU, Hoeh WR: **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* 2009, **183**:1575–1589.

43. Breton S, Ghiselli F, Passamonti M, Milani L, Stewart DT, Hoeh WR: **Evidence for a fourteenth mtDNA-encoded protein in the female-transmitted mtDNA of marine Mussels (Bivalvia: Mytilidae).** *PLoS One* 2011, **6**:e19365.

44. Milani L, Ghiselli F, Guerra D, Breton S, Passamonti M: **A comparative analysis of mitochondrial ORFans: New clues on their origin and role in species with Doubly Uniparental Inheritance of mitochondria**. *Genome Biol Evol* 2013, **5**:1408–1434.

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

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

47. Boyle EE, Etter RJ: **Heteroplasmy in a deep-sea protobranch bivalve suggests an ancient origin of doubly uniparental inheritance of mitochondria in Bivalvia**. *Mar Biol* 2013, **160**:413–422.

48. Xue T, Chen M, Wang G, Han Z, Li J: **The complete F-type mitochondrial genome of Chinese freshwater mussel Anodonta euscaphys**. *Mitochondrial DNA* 2015, **26**:263–264.

49. Rombel IT, Sykes KF, Rayner S, Johnston SA: **ORF-FINDER: A vector for high-throughput gene identification**. *Gene* 2002, **282**:33–41.

50. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, Taly J-F, Notredame C: **T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension**. *Nucleic Acids Res* 2011, **39**:W13–W17.

51. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3**:418–26.

52. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0**. *Mol Biol Evol* 2013, **30**:2725–2729.

53. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics**. *Nucleic Acids Res* 2004, **32**(Web Server):W273–W279.

54. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A: **Protein identification and analysis tools on the ExPASy server**. In *The proteomics protocols handbook*. Edited by Walker JM. Humana Press; 2005:571–607.

55. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157**:105–132.

56. Käll L, Krogh A, Sonnhammer EL.: **A Combined Transmembrane Topology and Signal Peptide Prediction Method**. *J Mol Biol* 2004, **338**:1027–1036.

57. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**:847–848.

58. Hofmann K, Stoffel W: **TMbase-A database of membrane spanning proteins segments**. *Biol Chem Hoppe-Seyler* 1993, **374**:166.

59. Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A: **Prediction of membrane-protein topology from first principles**. *Proc Natl Acad Sci* 2008, **105**:7177–

7181.

60. Rost B, Yachdav G, Liu J: **The PredictProtein server.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W321–6.

61. Nielsen H, Engelbrecht J, Brunak S, Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites Cited by me**. *Protein Eng* 1997, **10**:1–6.

62. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nature Methods* 2011:785–786.

63. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic Acids Res* 2010, **38**(Database):D161–D166.

64. Soding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction**. *Nucleic Acids Res* 2005, **33**(Web Server):W244–W248.

65. Karpenahalli MR, Lupas AN, Söding J: **TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences.** *BMC Bioinformatics* 2007, **8**.

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

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

68. Pons J-L, Labesse G: **@TOME-2: a new pipeline for comparative modeling of protein-ligand complexes.** *Nucleic Acids Res* 2009, **37**:W485–91.

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

70. Pesole G, Gissi C, De Chirico A, Saccone C: **Nucleotide substitution rate of mammalian mitochondrial genomes.** *J Mol Evol* 1999, **48**:427–434.

71. Djian P: **Evolution of simple repeats in DNA and their relation to human disease.** *Cell* 1998, **94**:155–60.

72. Björklund ÅK, Ekman D, Elofsson A: **Expansion of Protein Domain Repeats**. *PLoS Comput Biol* 2006, **2**:e114.

73. Manna S: **An overview of pentatricopeptide repeat proteins and their applications**. *Biochimie* 2015, **113**:93–99.

74. Sheprdson SP, Heard WH, Breton S, Hoeh WR: **Light and Transmission Electron Microscopy of Two Spermatogenic Pathways and Unimorphic Spermatozoa in Venustaconcha ellipsiformis (Conrad, 1836) (Bivalvia: Unionoida).** *Malacologia* 2012, **55**:263−284.

75. Terry LJ, Wente SR: **Flexible Gates: Dynamic Topologies and Functions for FG Nucleoporins in Nucleocytoplasmic Transport**. *Eukaryot Cell* 2009, **8**:1814–1827.

76. Schmitz-Linneweber C, Williams-Carrier RE, Williams-Voelker PM, Kroeger TS, Vichas A, Barkan A: **A pentatricopeptide repeat protein facilitates the trans-splicing of the maize chloroplast rps12 pre-mRNA.** *Plant Cell* 2006, **18**:2650–2663.

77. Ng F, Tang BL: **Pyruvate dehydrogenase complex (PDC) export from the mitochondrial matrix**. *Mol Membr Biol* 2014, **31**:207–210.

78. Bernardi P: **The mitochondrial permeability transition pore: a mystery solved?** *Front Physiol* 2013, **4**:1–12.

79. Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, Wan J, Kim S-J, Mehta H, Hevener AL, de Cabo R, Cohen P: **The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance**. *Cell Metab* 2015, **21**:443–454.

80. Tang BL: **Mitochondrial Protein in the Nucleus**. *CellBio* 2015, **4**:23–29.

81. Henderson IR, Navarro-Garcia F, Nataro JP: **The great escape: Structure and function of the autotransporter proteins**. *Trends in Microbiology* 1998:370–378.

82. Ulrich T, Oberhettinger P, Schütz M, Holzer K, Ramms AS, Linke D, Autenrieth IB, Rapaport D: **Evolutionary Conservation in Biogenesis of β-Barrel Proteins Allows Mitochondria to Assemble a Functional Bacterial Trimeric Autotransporter Protein**. *J Biol Chem* 2014, **289**:29457–29470.

83. Robert V, Volokhina EB, Senf F, Bos MP, Van Gelder P, Tommassen J: **Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif**. *PLoS Biol* 2006, **4**:1984–1995.

84. Gentle I: **The Omp85 family of proteins is essential for outer membrane biogenesis in mitochondria and bacteria**. *J Cell Biol* 2004, **164**:19–24.

85. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG: **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends Genet* 2009, **25**:404–413.

86. Kaessmann H: **Origins, evolution, and phenotypic impact of new genes**. *Genome Res* 2010, **20**:1313–1326.

87. Tautz D, Domazet-Lošo T: **The evolutionary origin of orphan genes.** *Nat Rev Genet* 2011, **12**:692–702.

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

89. Burzyn A, Wenne R: **Comparative Genomics of Marine Mussels (Mytilus spp .) Gender Associated mtDNA : Rapidly Evolving atp8**. *J Mol Evol* 2010, **71**:385–400.

90. Ghiselli F, Milani L, Guerra D, Chang PL, Breton S, Nuzhdin S V., Passamonti M: **Structure, Transcription, and Variability of Metazoan Mitochondrial Genome: Perspectives from an Unusual Mitochondrial Inheritance System**. *Genome Biol Evol* 2013,

5:1535–1554.

91. Boore JL, Brown WM: **Mitochondrial genomes of Galathealinum, Helobdella, and Platynereis: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa.** *Mol Biol Evol* 2000, **17**:87–106.

92. Burger G, Lang BF, Braun H, Marx S: **The enigmatic mitochondrial ORF ymf39 codes for ATP synthase chain b**. *Nucleic Acids Res* 2003, **31**:2353–2360.

93. Milani L: **Mitochondrial membrane potential : a trait involved in organelle inheritance ?** *Biol Lett* 2015, **11**:20150732.

94. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J, Mayrose I, Ming R, Perrin N, Ross L, Valenzuela N, Vamosi JC: **Sex Determination: Why So Many Ways of Doing It?** *PLoS Biol* 2014, **12**:e1001899.

95. Schwarzländer M, Finkemeier I: **Mitochondrial Energy and Redox Signaling in Plants**. *Antioxid Redox Signal* 2013, **18**:2122–2144.

96. Yu F, Shi J, Zhou J, Gu J, Chen Q, Li J, Cheng W, Mao D, Tian L: **ANK6, a mitochondrial ankyrin repeat protein, is required for male-female gamete recognition in Arabidopsis thaliana**. *PNAS* 2010, **107**:22332–22337.

97. Venetis C, Theologidis I, Zouros E, Rodakis GC: **No evidence for presence of maternal mitochondrial DNA in the sperm of Mytilus galloprovincialis males.** *Proc Biol Sci* 2006, **273**:2483–9.

98. Yusa Y, Breton S, Hoeh WR: **Population genetics of sex determination in Mytilus mussels: Reanalyses and a model**. *J Hered* 2013, **104**:380–385.

99. Henley WF, Neves RJ, Caceci T, Saacke RG: **Anatomical descriptions and comparison of the reproductive tracts of Utterbackia imbecillis and Villosa iris (Bivalvia: Unionidae)**. *Invertebr Reprod Dev* 2007, **50**:1–12.

100. Perlman SJ, Hodson CN, Hamilton PT, Opit GP, Gowen BE: **Maternal transmission, sex ratio distortion, and mitochondria.** *Proc Natl Acad Sci U S A* 2015, **112**:1–7.

101. Stewart DT, Kenchington ER, Singh RK, Zonros E: **Degree of selective constraint as an explanation of the different rates of evolution of gender-specific mitochondrial DNA lineages in the mussel Mytilus**. *Genetics* 1996, **143**:1349–1357.

102. Everett EM, Williams PJ, Gibson G, Stewart DT: **Mitochondrial DNA polymorphisms and sperm motility in Mytilus edulis (Bivalvia: Mytilidae)**. *J Exp Zool Part A Comp Exp Biol* 2004, **301A**:906–910.

103. Jha M, Côté J, Hoeh WR, Blier PU, Stewart DT: **Sperm motility in Mytilus edulis in relation to mitochondrial DNA polymorphisms: implications for the evolution of doubly uniparental inheritance in bivalves.** *Evolution (N Y)* 2008, **62**:99–106.

104. Chou JY, Leu JY: **The Red Queen in mitochondria: Cyto-nuclear co-evolution, hybrid breakdown and human disease**. *Front Genet* 2015, **6**.

105. Heath DJ: **Simultaneous hermaphroditism; cost and benefit**. *J Theor Biol* 1977, **64**:363–373.

106. Heath DJ: **Brooding and the evolution of hermaphroditism.** *J Theor Biol* 1979, **81**:151–155.

107. Ghiselin MT: **Sexual selection in hermaphrodites: Where did our ideas come from?** *Integr Comp Biol* 2006, **46**:368–372.

108. Ghiselin MT: **The evolution of hermphroditism among animals**. *Q Rev Biol* 1969, **44**:189–208.

109. Galbraith HS, Vaughn CC: **Effects of Reservoir Management on Abundance , Condition , Parasitism and Reproductive Traits of Downstream Mussels**. *River Res Appl* 2011, **27**:193–201.

110. Aldridge DC: **The morphology, growth and reproduction of unionidae (bivalvia) in a fenland waterwa y**. *J Molluscan Stud* 1999, **65**:47–60.

111. Viklund H, Elofsson A: **Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.** *Protein Sci* 2004, **13**:1908–1917.

112. Viklund H, Elofsson A: **OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar**. *Bioinformatics* 2008, **24**:1662–1668.

113. Granseth E, Viklund H, Elofsson A: **ZPRED: Predicting the distance to the membrane center for residues in α-helical membrane proteins**. In *Bioinformatics*. *Volume 22*; 2006:191–196.

114. Sigrist CJ a, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors.** *Brief Bioinform* 2002, **3**:265–274.

Appendice 1 : L'approche bioinformatique

Il peut être laborieux de déterminer la fonction d'une nouvelle protéine en laboratoire à l'aide de techniques biochimiques. Cependant, on peut prédire des caractéristiques structurales et fonctionnelles à partir de séquences protéiques en employant des méthodes de prédiction automatisées [111]. Une suite de programmes bioinformatiques peut être employée pour prédire des peptides signaux, des hélices transmembranaires, des domaines fonctionnels, et pour trouver des protéines connues avec des séquences ou des structures similaires pour nous informer sur la fonction potentielle de gènes nouvellement découverts comme ceux retrouvés dans les génomes mitochondriaux chez les moules d'eau douce.

Prédiction des hélices transmembranaires : Phobius est un programme qui prédit les peptides signaux et les hélices transmembranaires. Les programmes qui prédisent les hélices transmembranaires seulement peuvent parfois être induits en erreur par la présence de peptides signaux contenant des hélices alpha dans les séquences protéiques [56]. Pour résoudre ce problème, Phobius emploie un modèle de Markov caché (HMM, *Hidden Markov Model*) pour prédire ces deux structures en même temps, ce qui sépare les hélices alpha des peptides signaux des hélices transmembranaires, et facilite l'identification de l'orientation des hélices transmembranaires dans la membrane. Il est fiable pour des protéines qui contiennent les deux structures, mais conservateur si la protéine contient un peptide signal seulement. Phobius est parmi les meilleurs programmes permettant la détection des hélices transmembranaires, et, en combinaison avec le programme TMHMM (« *Transmembrane Hidden Markov Model* »), un outil intégré à InterProScan, on peut s'attendre à un taux d'erreur très faible [56].

Les hélices transmembranaires présentent une grande diversité structurale, et donc les méthodes de prédiction sont nombreuses et diverses. InterProScan compare la séquence protéique à l'étude à toutes les protéines connues et annotées dans toutes les bases de données qui sont membres de InterPro. Il intègre quatorze outils différents pour reconnaitre des signatures protéiques telles que les peptides signaux et hélices transmembranaires et donne un résultat visuel illustrant les prédictions de chaque outil (les 14 outils sont BlastProDom, HMMTigr, SignalPHMM, FPrintScan, ProfileScan, TMHMM, HMMPIR, HAMAP, HMMPanther, HMMPfam, PatternScan, Gene3D, HMMSmart et SuperFamily) [57]. TMPred

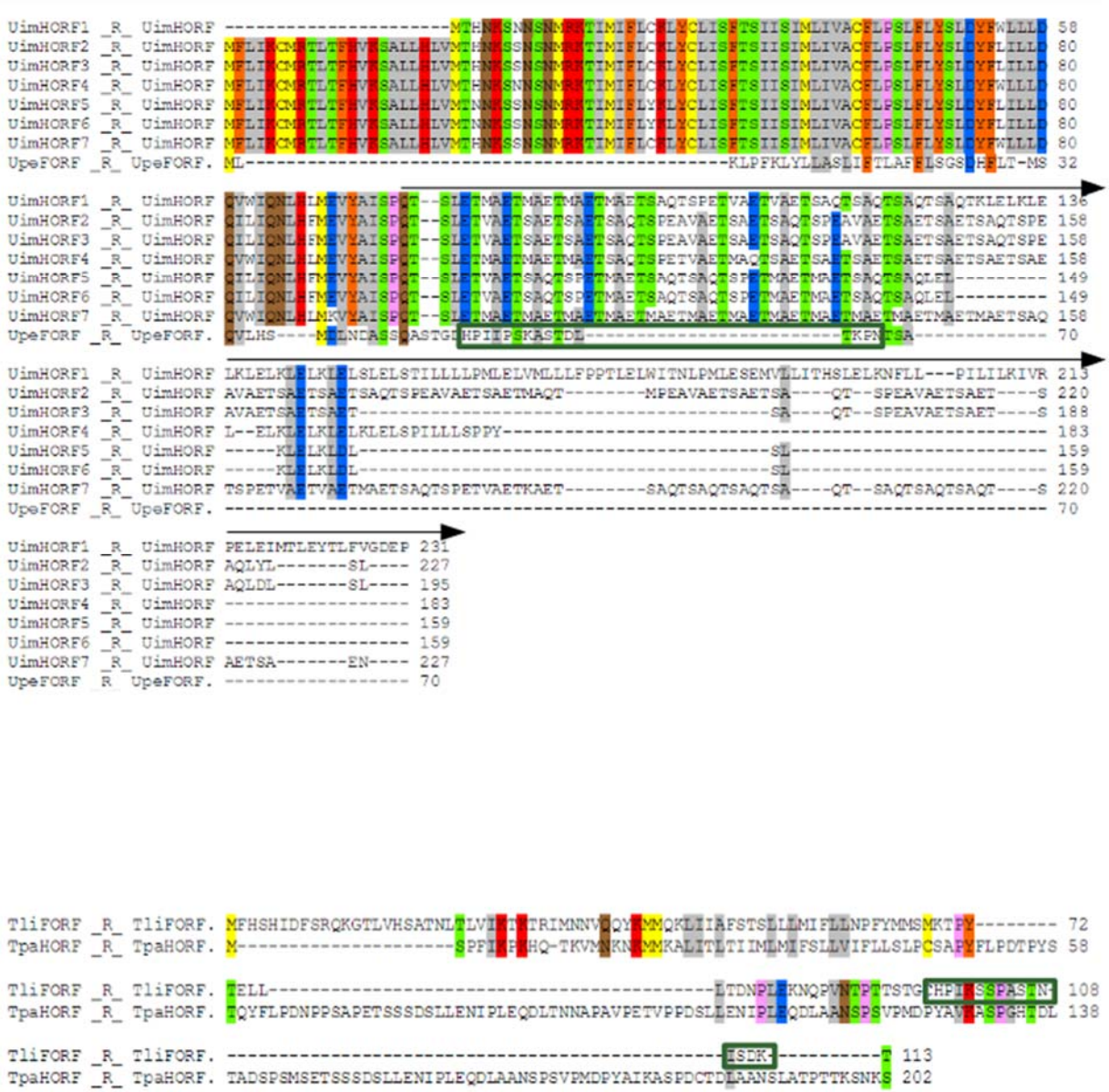
utilise une matrice de poids optimale pour comparer la séquence aux protéines dans la base de données TMbase. Il émet un score pour chaque résidu – un score de 500 ou plus est considéré significatif, et sera utilisé pour identifier une hélice transmembranaire et prédire son orientation dans la membrane [58]. Finalement, TOPCONS intègre cinq outils pour prédire des hélices transmembranaires : par exemple un qui aligne les séquences avec des modèles de protéines membranaires et un qui distingue les régions qui entrent dans, mais ne traversent pas la membrane. Puisque ces régions peuvent être mal identifiées comme transmembranaires, cette distinction est essentielle pour bien identifier l'orientation de la protéine dans la membrane. Les autres composants de TOPCONS [59, 112, 113] sont conçus pour imiter un translocon – c'est-à-dire qu'ils considèrent les caractéristiques physiques des résidus pour prédire comment ils interagissent avec la membrane, le milieu cellulaire, et d'autres résidus.

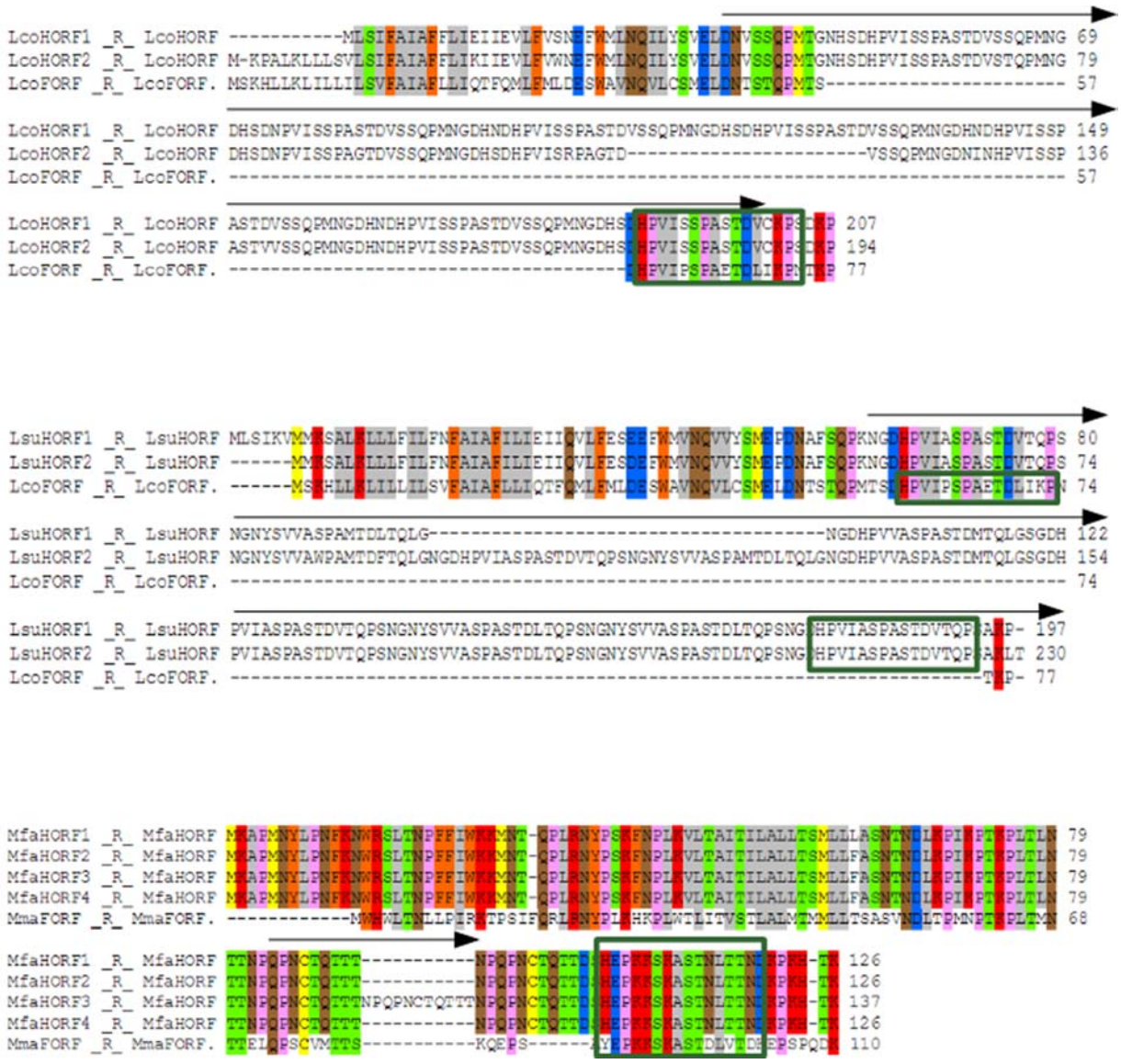
Prédiction de peptides signaux : PrediSi est parmi les programmes les plus avancés pour prédire des peptides signaux. Il emploie un réseau de neurones pour calculer trois scores : le *S-score*, qui indique la probabilité qu'un résidu fasse partie d'un peptide signal, le *C-score*, qui indique la probabilité qu'un résidu soit le premier acide aminé de la protéine mature, et le *Y-score*, qui combine les deux. Un *Y-score* élevé indique la présence d'un peptide signal. Si le *S-score* moyen de tous les résidus avant la position du *Y-score* maximal est >0.5, il prédit un peptide signal. Si les trois scores sont faibles, il est probable que la protéine n'est pas sécrétée [61]. Cette méthode est rapide et optimale pour des séquences de 60-100 acides aminés. Le programme SignalP emploie un autre réseau de neurones pour prédire des peptides signaux. Il est sensible et très précis, mais donne beaucoup de résultats faussement positifs. Il est généralement utilisé pour confirmer les résultats d'autres programmes [57]. Phobius et InterProScan (décrits ci-dessus) identifient également des peptides signaux putatifs.

Prédiction des domaines fonctionnels et de la fonction : Il existe différents programmes bioinformatiques pour prédire la fonction d'une protéine à partir de sa séquence en acides aminés. Par exemple, BLAST compare les séquences protéiques aux séquences de protéines connues dans les bases de données GenBank. La séquence en acides aminés est comparée, position par position, aux protéines connues pour trouver des homologues. La structure et la fonction de ces homologues peuvent nous informer sur la structure et la fonction de la protéine nouvellement découverte [66].

Le programme @tome-2 intègre 23 différents outils pour chercher des séquences homologues, prédire la structure de la protéine, et reconnaitre des repliements putatifs [68]. Le programme I-TASSER prédit des modèles tridimensionnels pour la séquence d'intérêt et les compare aux protéines dont la structure et la fonction sont connues [69]. Le programme HHpred analyse la séquence et l'aligne avec des protéines connues, et prédit les structures secondaires et tertiaires [64]. TPRpred est un programme similaire qui cherche exclusivement des répétitions de type tetratricopeptide, pentatricopeptide et SEL 1-like [65]. Motif Scan cherche des motifs (comme β-α-β) dans la base de données PROSITE et retourne plusieurs catégories de signifiance [114]. Finalement, le programme PredictProtein est un outil à usages multiples qui donne des résultats BLAST (alignements), cherche des motifs dans la base de données PROSITE, identifie des signaux de localisation nucléaire, des régions de faible complexité ou sans structure régulière, et prédit la structure secondaire, l'accessibilité aux solvants, les régions globulaires, les hélices transmembranaires, les domaines superhélices, les ponts disulfures, la localisation subcellulaire et les annotations/domaines fonctionnels [60]

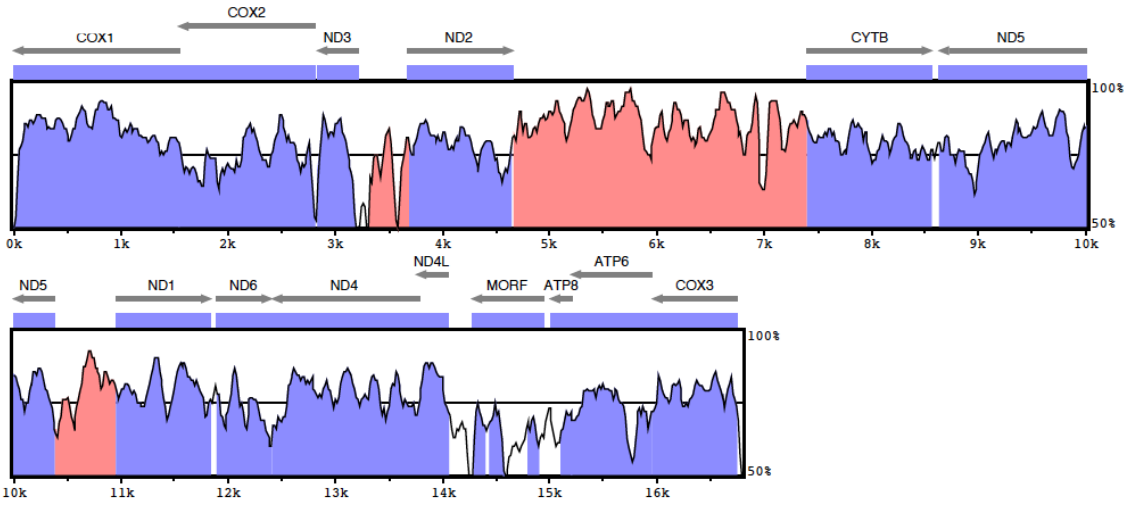
Appendice 2 : Figures supplémentaires



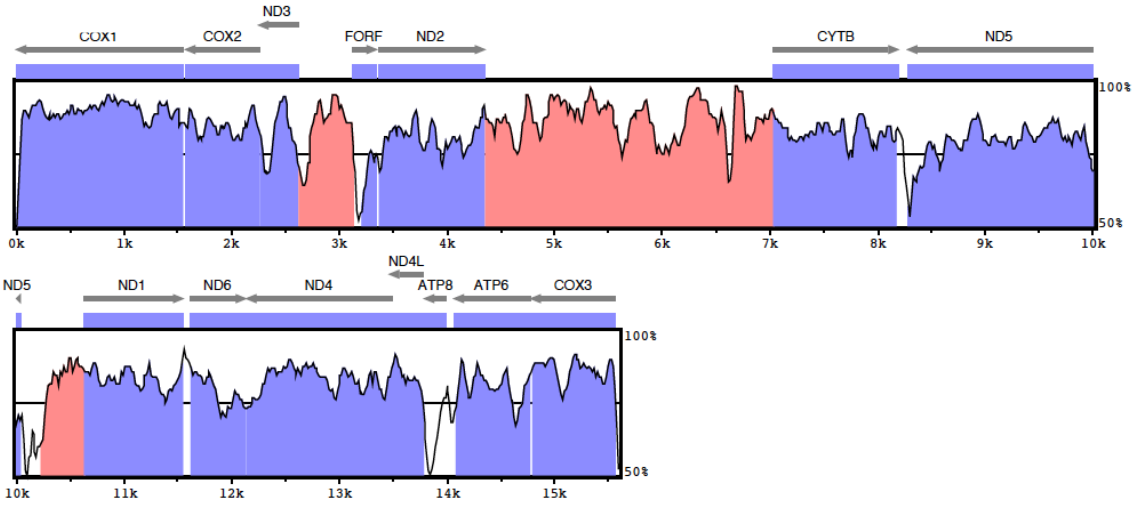


Supplementary Figure 1. Alignments of F-*ORFs* and H-*ORFs* of closely related species. Colour coding is applied to amino acid groups conserved in ≥70% of sequences. Grey, aliphatic amino acids; orange, aromatic amino acids; yellow, sulfur amino acids; green, amino acids bearing a hydroxyl group; red, basic amino acids; blue, acidic amino acids; brown, amino acids with an amide group; pink, cyclic amino acids. Green box: conserved C-terminal domain; blue underlining: repetitive sequences. UpeFORF, *U. peninsularis* F-*ORF*; UimHORF, *U. imbecillis* H-*ORF*; TliFORF, *T. lividus* F-*ORF*; TpaHORF, *T. parvum* H-*ORF*; MmaFORF, *M. margaritifera* F-*ORF*; MfaHORF, *M. falcata* H-*ORF*; LcoFORF, *L. complanata* F-*ORF*; LcoHORF, *L. compressa* H-*ORF*; LsuHORF, *L. subviridus* H-*ORF*.

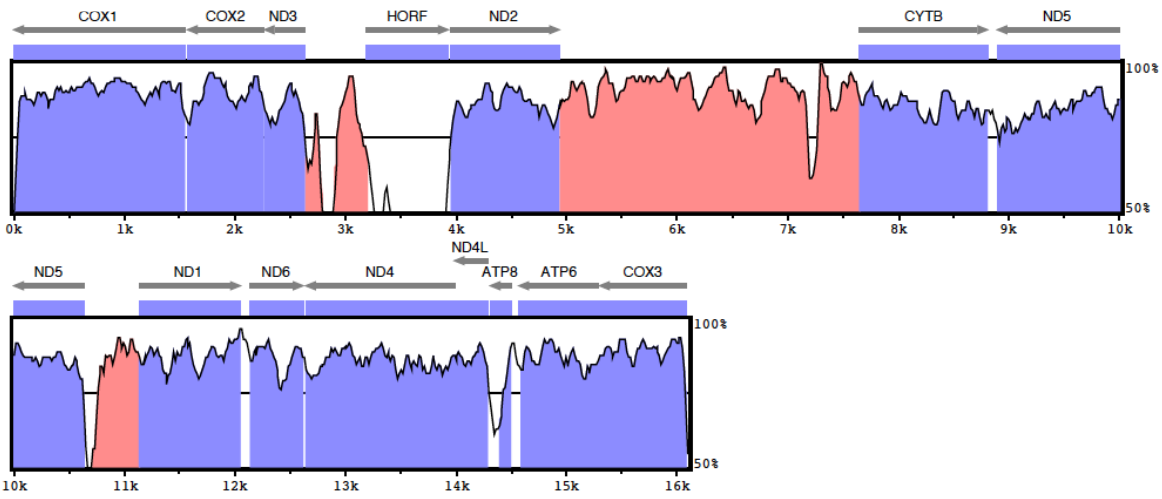
A. M vs M complete mitochondrial genomes



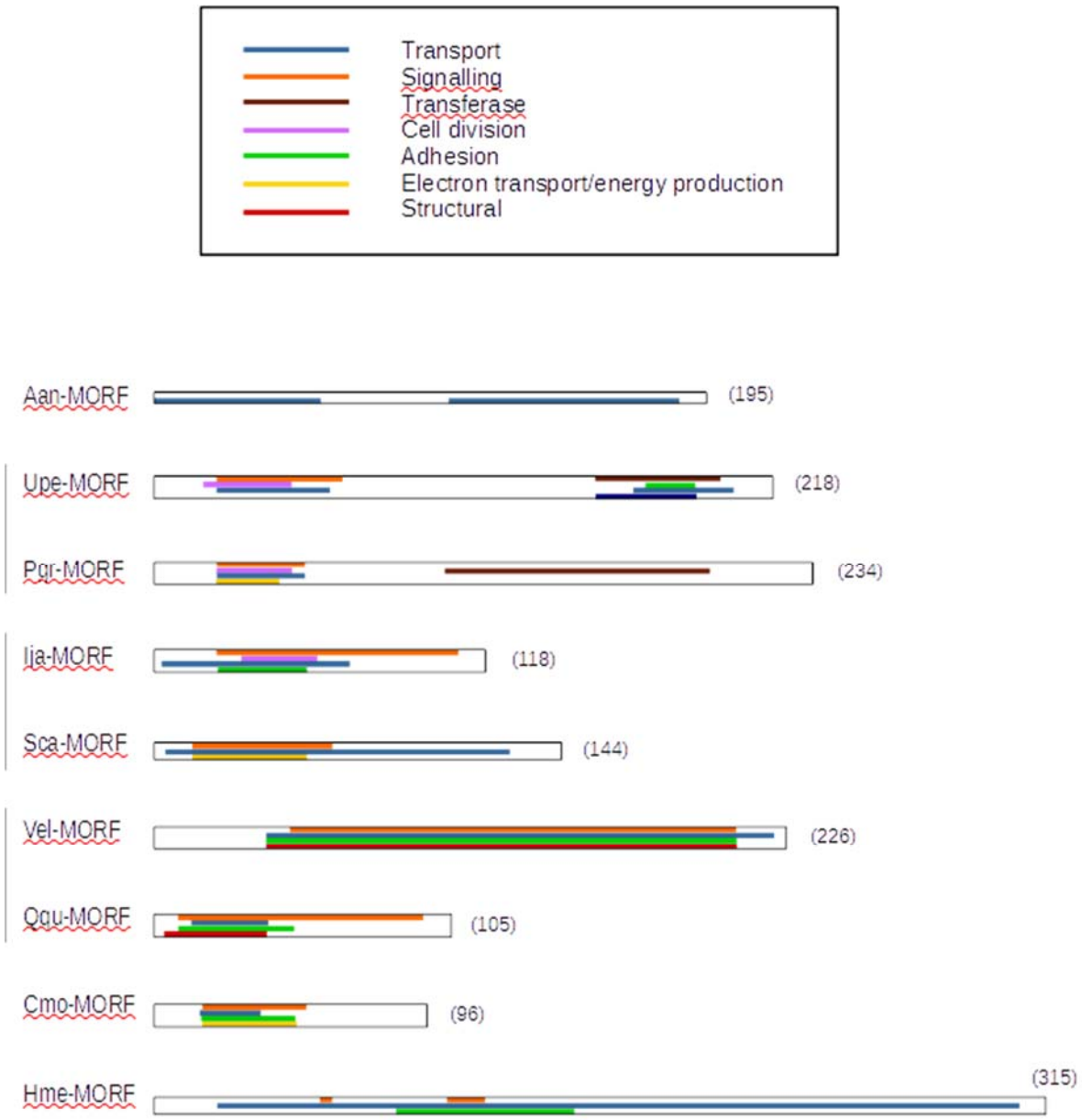
B. F vs F complete mitochondrial genomes

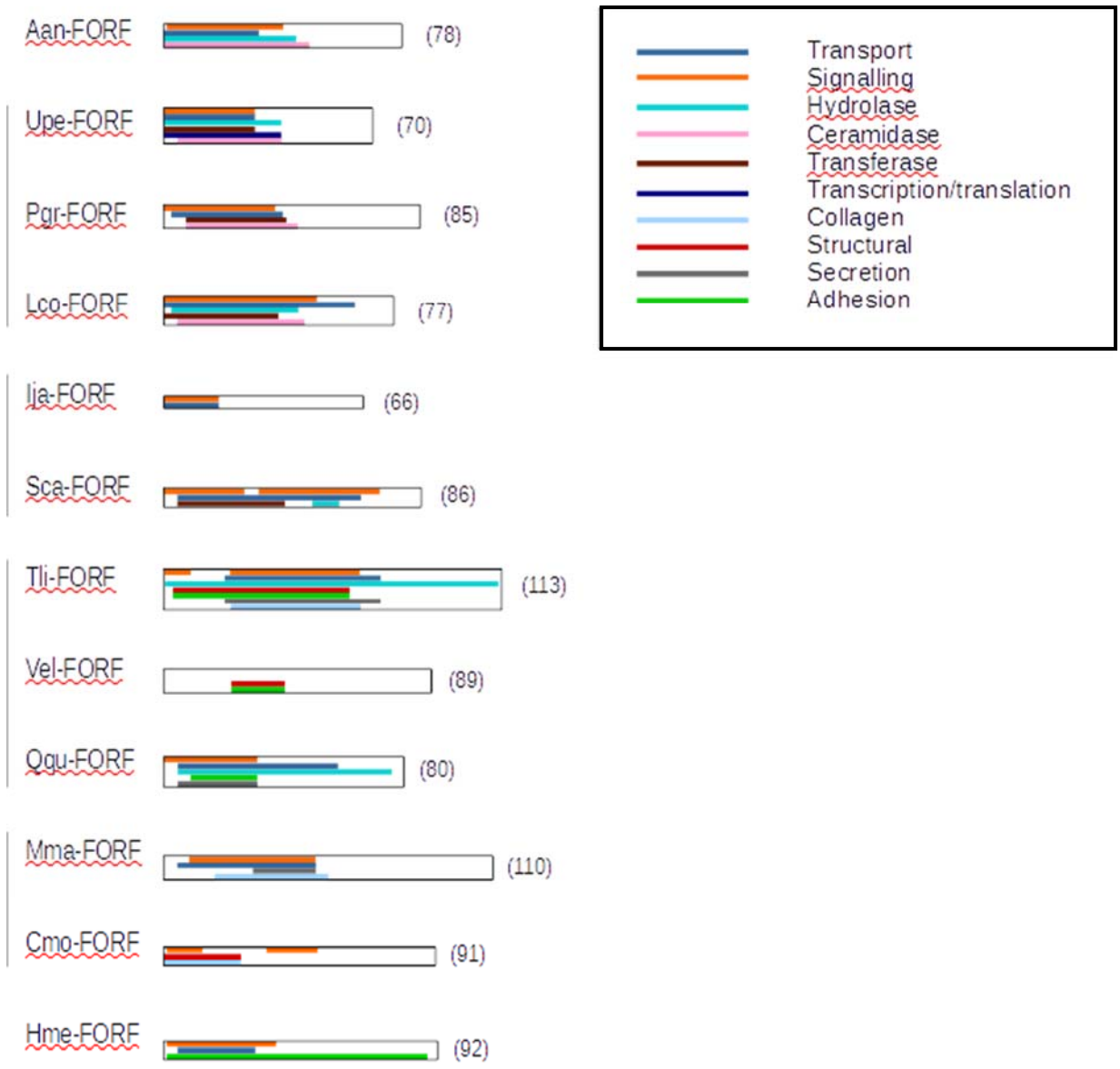


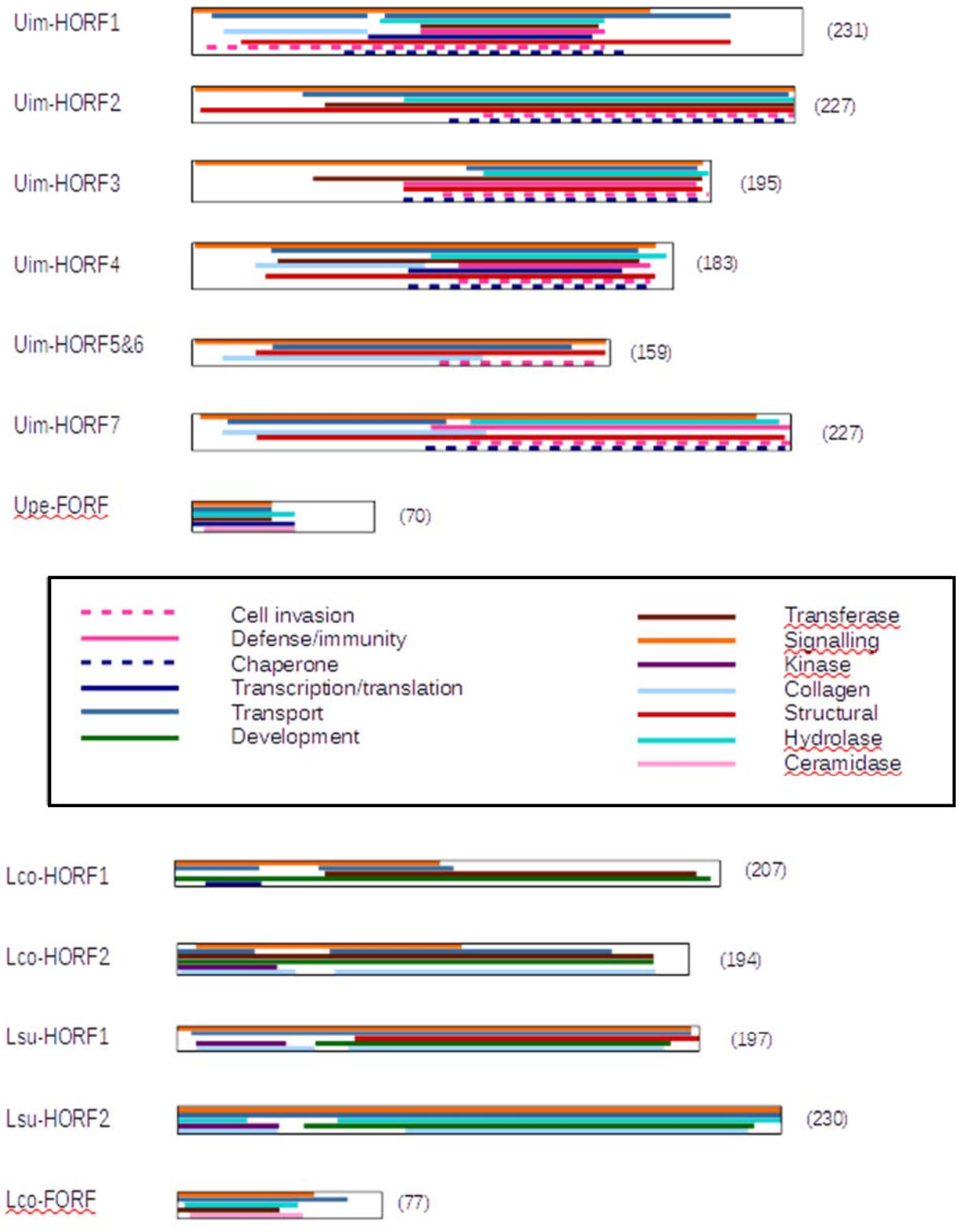
C. F vs H complete mitochondrial genomes

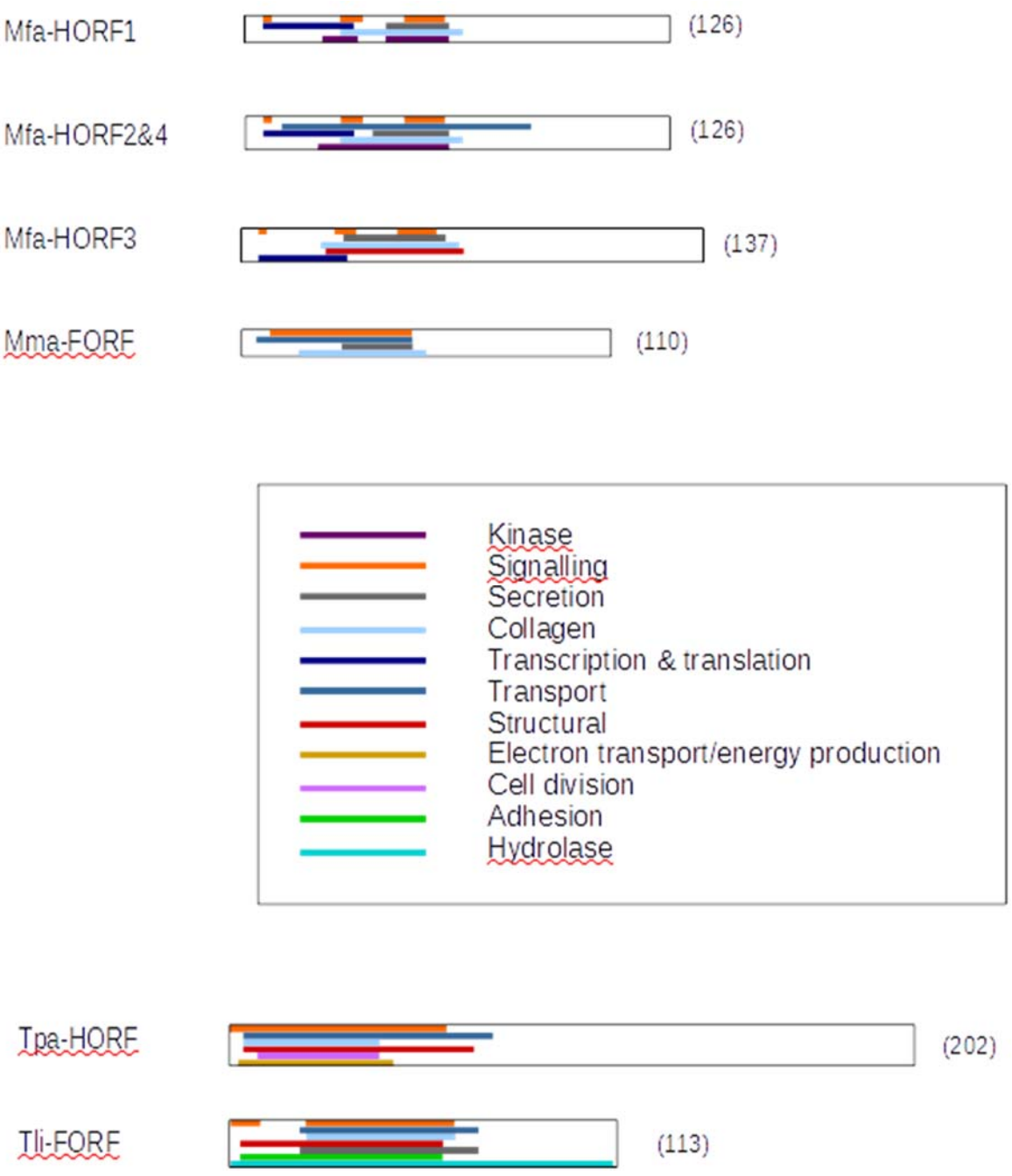


Supplementary Figure 2. Alignments of complete mitochondrial genomes of freshwater mussels with DUI. (A) M vs. M genome comparison between two closely related species (*Utterbackia peninsularis* and *Pyganodon grandis*, GenBank accession numbers HM856635 and FJ809754, respectively) showing that the M-*ORF* gene shows low level of sequence conservation compare to other protein-coding genes. (B) F vs. F genome comparison between two closely related species (*U. peninsularis* and *P. grandis*, GenBank accession numbers HM856636 and FJ809755, respectively) showing that the F-*ORF* gene shows low level of sequence conservation compare to other protein-coding genes. (C) F vs. H genome comparison between two closely related species (*Utterbackia peninsularis* and *U. imbecillis*, GenBank accession numbers HM856636 and HM856637, respectively) showing that the F-*ORF*/H-*ORF* gene region shows low level of sequence conservation compared to other protein-coding genes. Each graph shows the percent of conservation between genomes at any given coordinate. The top and bottom percentage bounds are shown to the right of every row. The pink regions are conserved non-protein-coding sequences, the dark blue regions are protein-coding genes.


Transport
Signalling
Transferase
Cell division
Adhesion
Electron transport/energy production
Structural
Aan-MORF
(195)
Upe-MORF
(218)
Pgr-MORF
(234)
Ija-MORF
(118)
Sca-MORF
(144)
Vel-MORF
(226)
Qgu-MORF
(105)
Cmo-MORF
(96)
Hme-MORF
(315)


Aan-FORF
(78)
Upe-FORF
(70)
Pgr-FORF
(85)
Lco-FORF
(77)
Ija-FORF
(66)
Sca-FORF
(86)
Tli-FORF
(113)
Vel-FORF
(89)
Qgu-FORF
(80)
Mma-FORF
(110)
Cmo-FORF
(91)
Hme-FORF
(92)
Transport
Signalling
Hydrolase
Ceramidase
Transferase
Transcription/translation
Collagen
Structural
Secretion
Adhesion


Uim-HORF1
(231)
Uim-HORF2
(227)
Uim-HORF3
(195)
Uim-HORF4
(183)
Uim-HORF5&6
(159)
Uim-HORF7
(227)
Upe-FORF
(70)
Cell invasion
Defense/immunity
Chaperone
Transcription/translation
Transport
Development
Transferase
Signalling
Kinase
Collagen
Structural
Hydrolase
Ceramidase
Lco-HORF1
(207)
Lco-HORF2
(194)
Lsu-HORF1
(197)
Lsu-HORF2
(230)
Lco-FORF
(77)



Supplementary Figure 3. Position of frequently recurring functions in HHpred and BLAST hits for (a) M-*ORFs*, (b) F-*ORFs*, and (c) and (d) H-*ORFs*. Hits with positions were grouped into categories and traced together, showing hot spots of functionality. Protein length in amino acids is indicated in parentheses. Subfamilies are indicated.

Appendice 3 : Tableaux supplémentaires

Supplementary Table I. Predicted transmembrane (TM) helices in M-*ORFs* and F-*ORFs.*

	TM Helices											
M-*ORF*	*Aan*	*Upe*	*Pgr*	*Lco*	*Ija*	*Sca*	*Tli*	*Vel*	*Qqu*	*Mma*	*Cmo*	*Hme*
Phobius	20-44	20-45	20-46		*21-42*	23-41		20-38	6-34		20-37	*20-42,* 54-77, *89-109*
InterProScan (TMHMM)	24-46	20-42	22-44		21-43	21-43		20-42	5-27		15-37	13-35, 55-77, 90-109
TMPred	23-41	21-38	*24-45*		*24-41*	*23-40*		*21-39*	7-27		*16-34*	*20-36* 54-73 *90-112*
TOPCONS	*24-44*	*18-38*	*22-42*		25-45	24-44		*15-35*	17-37		17-37	2-22, *69-89*
Predict Protein	26-43	22-39	24-44		19-42	22-39		21-38	17-32		17-33	21-38
Consensus	~23-44	~20-38	~22-44		~24-42	~22-41		~20-38	~10-30		~17-35	~19-34, 54-72, 90-110

F-*ORF*												
Phobius	-	-	-	-	-	-	*45-65*	*21-42*	*12-30*	31-53	-	-
InterProScan (TMHMM)	9-31	7-29	16-38	5-27	-	7-26	45-67	21-43	12-24	31-53	-	15-37
TMPred	*9-27*	*6-25*	*16-40*	*8-26*	*1-18*	*7-23*	45-68	*21-42*	12-30	*32-49*	*2-18*	*18-37*
TOPCONS	*9-29*	*8-28*	*16-36*	*8-28*	*2-22*	*6-26*	*41-61*	*21-41*	*10-30*	*31-51*	2-22	*17-37*
Predict Protein	9-26	8-25	14-31	8-25	1-18	8-25	44-66	20-42	16-33	32-49	1-18	17-31
Consensus	*~9-28*	*~7-27*	*~16-35*	*~8-26*	*~1-19*	*~7-25*	~45-66	*~21-42*	~12-29	~31-51	~2-19	*~17-36*

NOTE – All structures listed here were statistically supported by the programs used (Phobius posterior label probability>0.5; PrediSi score>0.5; SignalP score>D-cutoff 0.5; TMpred score>500; significance test not provided by the other programs). Numbers in italics represent TMHs predicted to be oriented from inside to outside, those underlined represent TMHs predicted to be oriented from outside to inside.

Supplementary Table II. Predicted signal peptides in M-*ORFs* and F-*ORFs.*

Signal Peptides												
M-*ORF*	*Aan*	*Upe*	*Pgr*	*Lco*	*Ija*	*Sca*	*Tli*	*Vel*	*Qqu*	*Mma*	*Cmo*	*Hme*
Phobius	-	-	-		-	-		-	-		-	-
InterProScan		-	-		-	-		-	-		-	-
PrediSi	CP43	CP 42*	CP 44		CP 40*	CP 35		CP 40*	CP 29*		CP 34	CP38*
SignalP	1-20	1-10	1-10		1-40	1-16		1-40	1-10		1-10	1-37
Consensus	-	-	-		1-40	-		1-40	-		-	1-38
F-*ORF*												
Phobius	1-26*	1-25*	1-33*	1-37*	1-26*	1-32*	-	-	-	-	1-20*	1-40*
InterProScan	-	-	-	-	-	-	-	-	-	-	-	-
PrediSi	CP26*	CP 25*	CP 33*	CP 25*	CP 17*	CP 32*	CP67	CP44	CP 32*	CP 51	CP 20*	CP 40*
SignalP	1-26*	1-19	1-36	1-37	1-20*	1-32*	1-18	1-44	1-32	1-51	1-20*	1-40
Consensus	1-26	~1-25	1-34	~1-33	~1-23	1-32	-	1-44	1-32	1-51	1-20	1-40

NOTE – All structures marked by an asterisk were statistically supported by the programs used. Those not marked with an asterisk were not statistically supported, but were predicted by multiple programs. (Phobius posterior label probability>0.5; PrediSi score>0.5; SignalP score>D-cutoff 0.5; TMpred score>500; significance test not provided by the other programs).

Supplementary Table III. Predicted transmembrane (TM) helices in H-*ORFs*.

TM Helix														
H-*ORF*	*Uim1*	*Uim2*	*Uim3*	*Uim4*	*Uim5&6*	*Uim7*	*Lsu1*	*Lsu2*	*Lco1*	*Lco2*	*Tpa*	*Mfa1*	*Mfa2&4*	*Mfa3*
Phobius	*21-46,* 52-73, *149-170,* 190-209	*37-61,* 67-84	*37-61*, 67-84	*44-68,* 74-95	*40-61,* 67-84	*44-68,* 74-94	*12-36*	-	-	*7-31*	-	-	-	-
InterProScan (TMHMM)	17-39	39-61	39-61	39-61	39-61	39-61	12-34	7-29	-	7-29	22-44	44-61	44-61	44-61
TMPred	23-50, *153-171*	54-72	54-72	*45-72*	*44-72*	*45-72*	*14-32*	*8-26*	*2-20*	*7-31*	*22-42*	*44-62*	*44-62*	*44-62*
TOPCONS	*150-170*, 189-209	*29-49*	-	52-72	*59-79*	*20-40,* 42-62	-	-	-	-	*22-42*	*44-64*	*44-64*	*44-64*
Predict Protein	22-41, 46-63, 195-212	41-65	43-67	46-64	51-65	42-66	16-33	11-29	1-18	10-28	26-44	43-61	43-61	43-60
Consensus	~22-46	~40-62	~44-65	~45-68	~45-65	~42-64	~13-33	~9-28	-	~7-30	~22-43	~44-62	~44-62	~44-62

NOTE – All structures listed here were statistically supported by the programs used (Phobius posterior label probability>0.5; PrediSi score>0.5; SignalP score>D-cutoff 0.5; TMpred score>500; significance test not provided by the other programs). Numbers in italics represent TMHs predicted to be oriented from inside to outside, those underlined represent TMHs predicted to be oriented from outside to inside.

Supplementary Table IV. Predicted signal peptides in H-*ORFs.*

Signal Peptides														
H-*ORF*	*Uim1*	*Uim2*	*Uim3*	*Uim4*	*Uim5&6*	*Uim7*	*Lsu1*	*Lsu2*	*Lco1*	*Lco2*	*Tpa*	*Mfa1*	*Mfa2&4*	*Mfa3*
Phobius	-	-	-	-	-	-	1-25*	1-19*	1-19*	-	1-47*	1-61*	1-61*	1-61*
InterProScan	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PrediSi	CP 168*	CP 69	CP 69	CP 69	CP 69	CP 69	CP 25*	CP 19*	CP 17	CP 18	CP 49*	CP 64*	CP 64*	CP 64*
SignalP	1-15	1-24	1-24	1-24	1-24	1-24	1-10	1-19	1-17	1-10	1-48	1-29	1-29	1-29
Consensus	-	-	-	-	-	-	1-25	1-19	1-18	-	~1-49	1-62	1-62	1-62

NOTE – All structures marked by an asterisk were statistically supported by the programs used. Those not marked with an asterisk were not statistically supported, but were predicted by multiple programs. (Phobius posterior label probability>0.5; PrediSi score>0.5; SignalP score>D-cutoff 0.5; TMpred score>500; significance test not provided by the other programs).

Supplementary Table V. Frequently recurring HHpred hits in F-*ORFs* and M-*ORFs*

F-*ORF* – probability (rank)	***Aan***	***Upe***	***Pgr***	***Lco***	***Ija***	***Sca***	***Tli***	***Vel***	***Qqu***	***Mma***	***Cmo***	***Hme***
Prepilin-type processing-associated H-X9-DG domain	99.31 (2)	99.34 (2)	99.27 (3)	99.32 (2)	99.23 (3)	99.37 (1)	99.23 (1)	99.30 (1)	99.37 (1)	99.11 (2)	99.04 (2)	99.25 (1)
Outer membrane insertion C-terminal signal	99.24 (3)	99.28 (3)	99.34 (2)	99.27 (3)	99.36 (2)	99.06 (2)	99.14 (2)	99.16 (2)	99.21 (2)	99.14 (1)	99.21 (1)	99.05 (3)
LPXTG cell wall anchor domain	99.47 (1)	99.47 (1)	99.45 (1)	99.46 (1)	99.44 (1)	98.91 (3)	98.81 (3)	98.97 (3)	99.02 (3)	98.87 (3)	98.83 (3)	99.10 (2)
X-X-X-Leu-X-X-Gly heptad repeats	98.03 (4)	98.08 (4)	97.97 (4)	98.05 (4)	97.91 (4)	97.69 (4)	97.99 (4)	97.97 (4)	97.98 (4)	97.75 (4)	97.70 (4)	97.66 (4)
GlyGly-CTERM domain	97.33 (5)	97.39 (5)	97.22(5)	97.32 (5)	97.58 (5)	96.90 (5)	97.08 (5)	97.29 (5)	97.38 (5)	97.15 (5)	96.86 (5)	96.93 (5)
Pentatricopeptide repeat domain	94.32 (6)	94.79 (6)	94.27 (7)	94.58 (6)	94.24 (6)	93.28 (6)	94.54 (6)	95.33 (6)	94.93 (6)	93.83 (6)	93.47 (6)	93.52 (6)

M-*ORF* – probability (rank)	***Aan***	***Upe***	***Pgr***	***Lco***	***Ija***	***Sca***	***Tli***	***Vel***	***Qqu***	***Mma***	***Cmo***	***Hme***
Prepilin-type processing-associated H-X9-DG domain	99.04 (2)	99.06 (2)	99.10 (1)		99.21 (2)	99.15 (1)		99.01 (2)	99.14 (2)		99.52 (1)	99.58 (1)
Outer membrane insertion C-terminal signal	99.24 (1)	96.16 (1)	98.89 (3)		99.25 (1)	98.75 (3)		99.11 (1)	99.19 (1)		99.20 (2)	99.19 (2)
LPXTG cell wall anchor domain	98.89 (3)	98.89 (3)	99.05 (2)		98.80 (3)	98.88 (2)		98.99 (3)	99.12 (3)		98.77 (3)	98.67 (3)
X-X-X-Leu-X-X-Gly heptad repeats	97.70 (4)	97.32 (4)	97.48 (4)		97.73 (4)	97.95 (4)		97.60 (4)	97.91 (4)		97.81 (4)	97.91 (4)
GlyGly-CTERM domain	97.26 (5)	97.24 (5)	97.15 (5)		96.74 (5)	97.04 (5)		96.47 (14)	97.57 (5)		97.14 (5)	96.67 (8)
Pentatricopeptide repeat domain	92.98 (6)	92.61 (6)	92.96 (6)		94.79 (6)	94.60 (6)		92.71 (48)	93.60 (6)		94.35 (6)	93.94 (47)

	Aan	*Upe*	*Pgr*	*Lco*	*Ija*	*Sca*	*Tli*	*Vel*	*Qqu*	*Mma*	*Cmo*	*Hme*
F-*ORF* – amino acid position												
Prepilin-type processing-associated H-X9-DG domain	19-22	18-21	26-29	18-29	13-15	2-9	34-36	11-13	1-4	44-49	48-50	80-85
Outer membrane insertion C-terminal signal	35-36	27-28	1-6	34-35	3-5	62-63	1-8	25-29	16-20	23-24	12-14	1-6
LPXTG cell wall anchor domain	55-60	47-52	62-67	54-59	1-15	4-22	95-96	19-34	10-25	32-48	72-76	8-35
X-X-X-Leu-X-X-Gly heptad repeats	47-54	39-46	54-61	46-53	57-65	4-7	18-22	71-78	62-69	49-56	18-27	8-12
GlyGly-CTERM domain	9-19	8-18	16-26	8-18	2-13	7-17	50-60	28-35	19-26	36-48	4-15	23-35
Pentatricopeptide repeat domain	26-46	31-38	46-53	38-45	14-18	16-23	30-49	7-25	63-66	16-23	59-68	60-69
	Aan	***Upe***	***Pgr***	***Lco***	***Ija***	***Sca***	***Tli***	***Vel***	***Qqu***	***Mma***	***Cmo***	***Hme***
M-*ORF* – amino acid position												
Prepilin-type processing-associated H-X9-DG domain	29-32	41-44	40-46		28-31	27-30		26-29	70-71		30-34	107-111
Outer membrane insertion C-terminal signal	57-64	53-60	55-59		40-44	6-7		158-164	19-21		51-56	53-56
LPXTG cell wall anchor domain	22-42	18-38	20-40		103-107	80-85		17-39	8-28		14-35	93-109
X-X-X-Leu-X-X-Gly heptad repeats	102-108	107-121	144-149		46-64	60-67		40-46	69-72		13-16	123-137
GlyGly-CTERM domain	30-42	26-38	28-40		22-35	21-36		22-35	6-18		16-26	98-109
Pentatricopeptide repeat domain	102-125	32-39	29-41		1-14	3-13		120-136	48-60		39-51	71-90

Supplementary Table VI. Frequently recurring HHpred hits in H-*ORFs*

	Uim1	***Uim2***	***Uim3***	***Uim4***	***Uim5&6***	***Uim7***	***Lsu1***	***Lsu2***	***Lco1***	***Lco2***	***Tpa***	***Mfa1***	***Mf2&4***	***Mfa3***
H-*ORF* – probability (rank)														
Prepilin-type processing-associated H-X9-DG domain	99.30 (1)	99.28 (1)	99.27 (2)	99.14 (2)	99.14 (2)	99.16 (2)	99.14 (2)	99.07 (1)	99.15 (1)	99.17 (1)	98.98 (2)	99.16 (2)	99.16 (2)	99.14 (2)
Outer membrane insertion C-terminal signal	99.27 (2)	99.22 (2)	99.33 (1)	99.28 (1)	99.19 (1)	99.28 (1)	99.24 (1)	98.86 (3)	98.77 (3)	98.82 (3)	98.46 (3)	99.23 (1)	99.23 (1)	99.19 (1)
LPXTG cell wall anchor domain	98.89 (3)	98.94 (3)	98.98 (3)	98.89 (3)	98.86 (3)	98.89 (3)	99.03 (3)	98.98 (2)	98.90 (2)	99.09 (2)	99.01 (1)	98.88 (3)	98.93 (3)	98.94 (3)
X-X-X-Leu-X-X-Gly heptad repeats	97.49 (4)	97.63 (7)	97.70 (21)	97.50 (11)	97.45 (6)	97.53 (4)	97.61 (4)	97.52 (4)	97.64 (4)	97.62 (4)	97.44 (4)	97.66 (4)	97.50 (4)	97.43 (4)
GlyGly-CTERM domain	96.92 (7)	97.09 (19)	97.05 (45)	96.95 (15)	96.72 (7)	96.97 (5)	96.98 (5)	96.88 (5)	96.49 (5)	97.02 (5)	96.69 (5)	96.99 (5)	96.99 (5)	96.95 (5)
Pentatricopeptide repeat domain	94.13 (23)	-	-	94.09 (30)	94.26 (8)	94.00 (8)	92.73 (8)	92.04 (8)	93.03 (6)	92.77 (8)	93.27 (6)	93.25 (6)	93.25 (6)	93.37 (6)
	Uim1	***Uim2***	***Uim3***	***Uim4***	***Uim5&6***	***Uim7***	***Lsu1***	***Lsu2***	***Lco1***	***Lco2***	***Tpa***	***Mfa1***	***Mf2&4***	***Mfa3***
H-*ORF* – amino acid position														
Prepilin-type processing-associated H-X9-DG domain	201-204	45-50	25-30	25-30	13-15	13-15	34-37	28-31	17-20	27-30	134-135	5-8	5-8	5-8
Outer membrane insertion C-terminal signal	49-52	75-81	71-74	71-74	71-74	71-74	1-5	91-93	23-24	33-34	40-44	5-7	5-7	5-7
LPXTG cell wall anchor domain	75-77	56-72	49-64	49-64	97-99	49-64	14-30	8-24	2-13	7-23	24-41	43-59	43-60	43-60
X-X-X-Leu-X-X-Gly heptad repeats	21-23	227-231	43-45	43-45	43-45	43-45	95-101	91-95	79-92	112-125	189-194	60-67	61-67	61-67

GlyGly-CTERM domain	49-57	58-71	71-79	71-79	55-70	96-97	14-24	8-18	2-12	7-17	33-44	47-59	47-59	47-60
Pentatricopeptide repeat domain	14-21	-	-	36-43	48-57	87-98	31-35	25-45	14-34	22-44	17-23	23-27	23-27	23-27

Supplementary Table VII. Hits to other motifs and domains in M-*ORFs* and F-*ORFs*

Motif or domain		***Aan***	***Upe***	***Pgr***	***Ija***	***Sca***	***Tli***	***Vel***	***Qqu***	***Cmo***	***Hme***
Lysine-rich region profile	M	X	X	X	X	X		X		X	X
	F										
Bipartite nuclear localization signal profile	M		X	X							X
	F										
RNA recognition motif in regulators of calcineurin and similar proteins	M		X								
	F										
Prokaryotic membrane lipoprotein lipid attachment site profile	M									X	
	F										
HIG1 domain family member	M	X	X								
	F										X
Telomerase reverse transcriptase (TEN domain)	M	X									
	F										
EGF-like-domain	M	X	X	X							
	F										X
Voltage-dependent anion channel	M										X
	F										
Histone H1-like protein Hc1	M										X
	F										
Microtubule-binding protein MIP-T3	M										X
	F										
Periplasmic protein TonB, links inner and outer membranes	M										X
	F										
Cell division protein FtsN	M										X
	F										
Plant ATP synthase F0	M									X	
	F										
DUF4381 Domain of unknown function	M									X	
	F										
E set domains	M									X	
	F										
Homeodomain-like	M									X	
	F										
PELOTA RNA binding domain	M		X								
	F										

Trigger factor ribosome-binding domain	M F		X								
DNaJ domain family member	M F				X						
Autophagy protein Apg6	M F							X			
Chromosome segregation ATPases	M F							X			
Chromosome segregation protein SMC, common bacterial type	M F							X			
TIGR03778 VPDSG-CTERM protein sorting domain	M F				X						
Bifunctional 2',3'-cyclic nucleotide 2'-phosphodiesterase/3'-nucleotidase precursor protein	M F					X					
TIGR04288 CGP-CTERM domain	M F									X	
Homodimeric domain of signal transducing histidine kinase	M F									X	
Virus attachment protein globular domain	M F									X	
Opacity-associated protein A N-terminal motif	M F						X	X	X		

NOTE – *Lco* and *Mma* M-*ORFs* and F-*ORFs* did not return any motifs or domains other than the frequently recurring HHpred hits.

Supplementary Table VIII. Hits to other motifs and domains in H-*ORFs*

Motif or domain	***Uim***						***Lsu***		***Mfa***		
	1	**2**	**3**	**4**	**5&6**	**7**	**1**	**2**	**1**	**2&4**	**3**
Response regulator receiver domain protein (CheY-like)	X										
Mitochondria Localisation Sequence		X	X								
ribonuclease E		X	X			X					
Ehrlichia tandem repeat		X	X			X					
Terminal organelle assembly protein TopJ		X	X								
Bifunctional 2',3'-cyclic nucleotide 2'-phosphodiesterase/3'-nucleotidase precursor protein				X							
TIGR03544 DivIVA domain									X	X	X
EGF-like-domain									X		X
Herpes virus major outer envelope glycoprotein (BLLF1)							X	X			

NOTE – *Lco* and *Tpa* did not return any motifs or domains other than the frequently recurring HHpred hits.

Supplementary Table IX. Filtered hmmsearch output for the M-*ORF* and F-*ORF* HMM profiles built using default parameters with hmmbuild.

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
M-*ORF*	UniProtKB	F4ZG80_9BIVA	M-specific morf protein	Eukaryota	Utterbackia peninsularis	1	1	**174.6**	**5.20E-48**
		V9PBU4_9BIVA	M-ORF	Eukaryota	Solenaia carinatus	1	1	**94.8**	**1.70E-23**
		A0A023I1E9_ANOAN	M-ORF	Eukaryota	Anodonta anatina	1	1	**88.8**	**1.20E-21**
		A0A023I1I6_ANOAN	M-ORF	Eukaryota	Anodonta anatina	1	1	**88.6**	**1.40E-21**
		A0A0F4GXW8_9PEZI	DUF221-domain-containing protein	Eukaryota	Zymoseptoria brevis	1	1	**30.5**	**0.001**
		G2RQY3_BACME	Excalibur domain protein	Bacteria	Bacillus megaterium WSH-002	1	1	**30.3**	0.0011
		A0A068N778_BACCE	Group-specific protein	Bacteria	Bacillus cereus	1	1	**29.0**	0.0029
		A0A0D0GUL4_BACTM	Bacillus thuringiensis serovar morrisoni strain HD 600 BG10.Contig244, whole genome shotgun sequence	Bacteria	Bacillus thuringiensis subsp. morrisoni	1	1	**28.2**	0.0048
		G3H659_CRIGR	CKLF-like MARVEL transmembrane domain-containing protein 2B	Eukaryota	Cricetulus griseus	1	1	**27.4**	0.0087
		K2G8H3_9BACT	RNA binding S1 protein	Bacteria	uncultured bacterium (gcode 4)	1	1	**27.4**	0.0089
		R7N780_9FIRM	Electron transport complex subunit E	Bacteria	Firmicutes bacterium CAG:95	1	0	**26.6**	0.016
		A0A0E0W0K4_BACAN	Group-specific protein	Bacteria	Bacillus anthracis str. H9401	1	0	**26.3**	0.019
		Q63BB7_BACCZ	Group-specific protein	Bacteria	Bacillus cereus (strain ZK / E33L)	1	0	**26.3**	0.019
		C1H9F4_PARBA	Nucleolar protein NOP56	unclassified sequences	unclassified	1	0	**25.9**	0.025
		A0A0D6M554_9BILA	SnoRNA binding domain protein	Eukaryota	Ancylostoma ceylanicum	1	0	**25.6**	0.031
		A5KSC4_9BACT	ATP synthase subunit b	Bacteria	candidate division TM7 genomosp. GTL1	1	0	**24.9**	0.052
		A0A061B2Y9_CYBFA	CYFA0S08e02300g1_1	Eukaryota	Cyberlindnera fabianii	1	0	**24.6**	0.061
		Q8EWL2_MYCPE	Putative uncharacterized protein MYPE1910	Bacteria	Mycoplasma penetrans (strain HF-2)	1	0	**24.4**	0.076
		A0A098DB54_GIBZA	Fusarium graminearum chromosome 1, complete genome	Eukaryota	Gibberella zeae	1	0	**24.3**	0.079

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		H2J4N9_MARPK	ATP synthase subunit b	Bacteria	Marinitoga piezophila (strain DSM 14283 / JCM 11233 / KA3)	1	0	**24.3**	0.081
M-*ORF*	UniProtKB	A0A061CBD1_LACDL	Hypothetical membrane protein	Bacteria	Lactobacillus delbrueckii subsp. lactis	1	0	**23.7**	0.12
		H2B2B1_KAZAF	KAFR0L01510 protein	Eukaryota	Kazachstania africana (strain ATCC 22294 / BCRC 22015 / CBS 2517 / CECT 1963 / NBRC 1671 / NRRL Y-8276)	1	0	**23.6**	0.12
		A0A023FMS2_9ACAR	Putative ribosome bioproteinsis protein	Eukaryota	Amblyomma cajennense	1	0	**23.4**	0.15
		S9UXZ4_9TRYP	Cellular retinaldehyde-binding protein/triple function domain-containing protein	Eukaryota	Strigomonas culicis	1	0	**23.3**	0.16
		C4L4A3_EXISA	Glycosyl transferase family 51	Bacteria	Exiguobacterium sp. (strain ATCC BAA-1283 / AT1b)	1	0	**23.1**	0.18
		A0A085C9C0_BACIU	Membrane protein	Bacteria	Bacillus subtilis	1	0	**22.9**	0.21
		YTTA_BACSU	Uncharacterized membrane protein YttA	Bacteria	Bacillus subtilis (strain 168)	1	0	**22.9**	0.21
		A6TVR5_ALKMQ	Integral membrane sensor signal transduction histidine kinase	Bacteria	Alkaliphilus metalliredigens (strain QYMF)	1	0	**22.8**	0.22
		Q28264_CANFA	Junctional sarcoplasmic reticulum protein	Eukaryota	Canis familiaris	1	0	**22.8**	0.23
		L7MMA2_OESDE	RIC-3	Eukaryota	Oesophagostomum dentatum	1	0	**22.4**	0.3
		C6H7B9_AJECH	Sec14 cytosolic factor	Eukaryota	Ajellomyces capsulatus (strain H143)	1	0	**22.3**	0.33
		L9VVP2_9EURY	ATPase AAA containing von Willebrand factor type A (VWA) domain-like protein	Archaea	Natronorubrum tibetense GA33	1	0	**22.2**	0.35
		B5RV46_DEBHA	Vacuolar protein sorting-associated protein 29	Eukaryota	Debaryomyces hansenii (strain ATCC 36239 / CBS 767 / JCM 1990 / NBRC 0083 / IGC 2968)	1	0	**22.0**	0.39
		K2HQK2_ENTNP	Major facilitator superfamily protein	Eukaryota	Entamoeba nuttalli (strain P19)	1	0	**21.7**	0.48
		W4GC94_9STRA	tRNA pseudouridine(55) synthase	Eukaryota	Aphanomyces astaci	1	0	**21.5**	0.57
		F0U682_AJEC8	SEC14 cytosolic factor	Eukaryota	Ajellomyces capsulatus (strain	1	0	**21.5**	0.58

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
					H88)				
		A0A078J4T8_BRANA	BnaCnng34340D protein	Eukaryota	Brassica napus	1	0	**21.4**	0.6
		I2H4F9_TETBL	TBLA0E02080 protein	Eukaryota	Tetrapisispora blattae (strain ATCC 34711 / CBS 6284 / DSM 70876 / NBRC 10599 / NRRL Y-10934 / UCD 77-7)	1	0	**21.4**	0.61
M-*ORF*	UniProtKB	H0DG78_9STAP	Nuclease-like protein	Bacteria	Staphylococcus pettenkoferi VCU012	1	0	**21.4**	0.62
		M3UR84_ENTHI	Major facilitator superfamily protein	Eukaryota	Entamoeba histolytica HM-1:IMSS-B	1	0	**21.1**	0.74
		K0NZP8_9LACO	Hypothetical membrane protein	Bacteria	Lactobacillus equicursoris DSM 19284 = JCM 14600 = CIP 110162	1	0	**21.1**	0.75
		A0A090BR33_KLUMA	Nucleolar protein 56	Eukaryota	Kluyveromyces marxianus	1	0	**21.1**	0.75
		W0TC29_KLUMA	Nucleolar protein 56	Eukaryota	Kluyveromyces marxianus DMKU3-1042	1	0	**21.1**	0.75
		C5D4P5_GEOSW	Penicillin-binding protein transpeptidase	Bacteria	Geobacillus sp. (strain WCH70)	1	0	**21.0**	0.8
		R6HQS4_9PROT	Putative uncharacterized membrane protein	Bacteria	Azospirillum sp. CAG:260	1	0	**21.0**	0.8
		F9N4N5_9FIRM	ATP synthase subunit b	Bacteria	Veillonella sp. oral taxon 780 str. F0422	1	0	**20.8**	0.96
		H2B1U5_KAZAF	KAFR0K02390 protein	Eukaryota	Kazachstania africana (strain ATCC 22294 / BCRC 22015 / CBS 2517 / CECT 1963 / NBRC 1671 / NRRL Y-8276)	1	0	**20.7**	0.97
	SwissProt	YTTA_BACSU	Uncharacterized membrane protein YttA	Bacteria	Bacillus subtilis (strain 168)	1	1	**22.9**	0.0023
		YZVL_CAEEL	Uncharacterized NOP5 family protein K07C5.4	Eukaryota	Caenorhabditis elegans	1	0	18.7	0.046
		PROQ_VIBF1	RNA chaperone ProQ	Bacteria	Vibrio fischeri (strain ATCC 700601 / ES114)	1	0	17.9	0.08
		PROQ_VIBFM	RNA chaperone ProQ	Bacteria	Vibrio fischeri (strain MJ11)	1	0	17.9	0.08
		HDAC1_CHICK	Histone deacetylase 1	Eukaryota	Gallus gallus	1	0	17.4	0.12

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		PROQ_PSYIN	RNA chaperone ProQ	Bacteria	Psychromonas ingrahamii (strain 37)	1	0	17.0	0.15
		YYAB_BACSU	Uncharacterized protein YyaB	Bacteria	Bacillus subtilis (strain 168)	1	0	16.9	0.17
		NU3M_YARLI	NADH-ubiquinone oxidoreductase chain 3	Eukaryota	Yarrowia lipolytica (strain CLIB 122 / E 150)	2	0	16.2	0.27
		LRC59_RAT	Leucine-rich repeat-containing protein 59	Eukaryota	Rattus norvegicus	1	0	16.1	0.28
		CT47A_HUMAN	Cancer/testis antigen 47A	Eukaryota	Homo sapiens	1	0	15.9	0.33
		PROQ_ALISL	RNA chaperone ProQ	Bacteria	Aliivibrio salmonicida LFI1238	1	0	15.8	0.34
		LRC59_HUMAN	Leucine-rich repeat-containing protein 59	Eukaryota	Homo sapiens	1	0	15.8	0.36
		TNSB_ECOLX	Transposon Tn7 transposition protein TnsB	Bacteria	Escherichia coli	1	0	15.7	0.37
M-*ORF*	SwissProt	RPN2_CANGA	26S proteasome regulatory subunit RPN2	Eukaryota	Candida glabrata (strain ATCC 2001 / CBS 138 / JCM 3761 / NBRC 0622 / NRRL Y-65)	1	0	15.5	0.44
		Y377_MYCGE	Uncharacterized protein MG377	Bacteria	Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195)	1	0	15.3	0.49
		ATPF_XYLFT	ATP synthase subunit b	Bacteria	Xylella fastidiosa (strain Temecula1 / ATCC 700964)	1	0	14.9	0.67
		ATPF_XYLF2	ATP synthase subunit b	Bacteria	Xylella fastidiosa (strain M23)	1	0	14.9	0.67
		ATPF_BURVG	ATP synthase subunit b	Bacteria	Burkholderia vietnamiensis (strain G4 / LMG 22486)	1	0	14.8	0.72
		ATPF_BURCC	ATP synthase subunit b	Bacteria	Burkholderia cenocepacia (strain MC0-3)	1	0	14.8	0.74
		ATPF_BURCA	ATP synthase subunit b	Bacteria	Burkholderia cenocepacia (strain AU 1054)	1	0	14.8	0.74
		ATPF_BURCH	ATP synthase subunit b	Bacteria	Burkholderia cenocepacia (strain HI2424)	1	0	14.8	0.74
		ATPF_BURCJ	ATP synthase subunit b	Bacteria	Burkholderia cenocepacia (strain ATCC BAA-245 / DSM 16553 /	1	0	14.8	0.74

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
					LMG 16656 / NCTC 13227 / J2315 / CF5610)				
		ATPF_BURM1	ATP synthase subunit b	Bacteria	Burkholderia multivorans	1	0	14.7	0.75
		ATPF_BURA4	ATP synthase subunit b	Bacteria	Burkholderia ambifaria (strain MC40-6)	1	0	14.6	0.83
		ATPF_BURCM	ATP synthase subunit b	Bacteria	Burkholderia ambifaria (strain ATCC BAA-244 / AMMD)	1	0	14.6	0.83
		SHDAG_HDVAM	Small delta antigen	Viruses	Hepatitis delta virus genotype I (isolate American)	2	0	14.6	0.84
		OTCC_TREDE	Ornithine carbamoyltransferase, catabolic	Bacteria	Treponema denticola (strain ATCC 35405 / CIP 103919 / DSM 14222)	1	0	14.4	0.93
	PDB	3x1l_C	Cmr4	Archaea	Archaeoglobus fulgidus DSM 4304	2	0	16.0	0.15
	PDB	1a92_A	DELTA ANTIGEN	Viruses	Hepatitis delta virus	1	0	14.8	0.36
	QfO	YTTA_BACSU	Uncharacterized membrane protein YttA	Bacteria	Bacillus subtilis (strain 168)	2	1	**22.9**	0.0032
		YZVL_CAEEL	Uncharacterized NOP5 family protein K07C5.4	Eukaryota	Caenorhabditis elegans	1	0	18.7	0.064
M-*ORF*	QfO	Q9LTV0_ARATH	NOP56-like pre RNA processing ribonucleoprotein	Eukaryota	Arabidopsis thaliana	1	0	18.4	0.078
		HDAC1_CHICK	Histone deacetylase 1	Eukaryota	Gallus gallus	1	0	17.4	0.16
		YYAB_BACSU	Uncharacterized protein YyaB	Bacteria	Bacillus subtilis (strain 168)	1	0	16.9	0.23
		Q7S9Y2_NEUCR	NMDA receptor-regulated protein 1	Eukaryota	Neurospora crassa (strain ATCC 24698 / 74-OR23-1A / CBS 708.71 / DSM 1257 / FGSC 987)	1	0	16.6	0.27
		Q9LJA1_ARATH	Expressed protein	Eukaryota	Arabidopsis thaliana	1	0	16.3	0.35
		NU3M_YARLI	NADH-ubiquinone oxidoreductase chain 3	Eukaryota	Yarrowia lipolytica (strain CLIB 122 / E 150)	2	0	16.2	0.38
		LRC59_RAT	Leucine-rich repeat-containing protein 59	Eukaryota	Rattus norvegicus	1	0	16.1	0.39
		CT47A_HUMAN	Cancer/testis antigen 47A	Eukaryota	Homo sapiens	1	0	15.9	0.46
		H2QDF0_PANTR	Leucine rich repeat containing 59	Eukaryota	Pan troglodytes	1	0	15.8	0.5
		LRC59_HUMAN	Leucine-rich repeat-containing protein 59	Eukaryota	Homo sapiens	1	0	15.8	0.5

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
	Pfamseq	G2RQY3_BACME	Excalibur domain protein	Bacteria	Bacillus megaterium WSH-002	1	1	**30.3**	0.0018
		G3H659_CRIGR	CKLF-like MARVEL transmembrane domain-containing protein 2B	Eukaryota	Cricetulus griseus	1	0	**27.4**	0.014
		K2G8H3_9BACT	RNA binding S1 protein	Bacteria	uncultured bacterium (gcode 4)	1	0	**27.4**	0.014
		R7N780_9FIRM	Electron transport complex subunit E	Bacteria	Firmicutes bacterium CAG:95	1	0	**26.6**	0.026
		J6PCC7_BACAN	Group-specific protein	Bacteria	Bacillus anthracis str. BF1	1	0	**26.3**	0.03
		W8HQB3_BACAN	Group-specific protein	Bacteria	Bacillus anthracis str. SVA11	1	0	**26.3**	0.03
		J6DVY5_BACAN	Group-specific protein	Bacteria	Bacillus anthracis str. UR-1	1	0	**26.3**	0.03
		I0D291_BACAN	Group-specific protein	Bacteria	Bacillus anthracis str. H9401	1	0	**26.3**	0.03
		Q63BB7_BACCZ	Group-specific protein	Bacteria	Bacillus cereus (strain ZK / E33L)	1	0	**26.3**	0.03
		C1H9F4_PARBA	Nucleolar protein NOP56	unclassified sequences	unclassified	1	0	**25.9**	0.041
		E4SY79_LACDN	Hypothetical membrane protein	Bacteria	Lactobacillus delbrueckii subsp. bulgaricus ND02	1	0	**25.8**	0.043
		A5KSC4_9BACT	ATP synthase subunit b	Bacteria	candidate division TM7 genomosp. GTL1	1	0	**24.9**	0.083
		Q8EWL2_MYCPE	Putative uncharacterized protein MYPE1910	Bacteria	Mycoplasma penetrans (strain HF-2)	1	0	**24.4**	0.12
M-*ORF*	Pfamseq	H2J4N9_MARPK	ATP synthase subunit b	Bacteria	Marinitoga piezophila (strain DSM 14283 / JCM 11233 / KA3)	1	0	**24.3**	0.13
		A0A023FMS2_9ACAR	Putative ribosome bioproteinsis protein	Eukaryota	Amblyomma cajennense	1	0	**23.4**	0.24
		S9UXZ4_9TRYP	Cellular retinaldehyde-binding protein/triple function domain-containing protein	Eukaryota	Strigomonas culicis	1	0	**23.3**	0.26
		A0A031LCL5_ENTFC	MAEBL family membrane protein	Bacteria	Enterococcus faecium VRE0576	1	0	**23.2**	0.28
		C4L4A3_EXISA	Glycosyl transferase family 51	Bacteria	Exiguobacterium sp. (strain ATCC BAA-1283 / AT1b)	1	0	**23.1**	0.29
		C0SHR5_PARBP	Nucleolar protein 5A	Eukaryota	Paracoccidioides brasiliensis (strain Pb03)	1	0	**23.0**	0.32

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		V5MXR1_BACIU	Putative membrane protein yttA	Bacteria	Bacillus subtilis PY79	1	0	**22.9**	0.34
		YTTA_BACSU	Uncharacterized membrane protein YttA	Bacteria	Bacillus subtilis (strain 168)	1	0	**22.9**	0.34
		M1U5M5_BACIU	YttA	Bacteria	Bacillus subtilis subsp. subtilis 6051-HGW	1	0	**22.9**	0.34
		J7JVR5_BACIU	YttA	Bacteria	Bacillus subtilis QB928	1	0	**22.9**	0.34
		A6TVR5_ALKMQ	Integral membrane sensor signal transduction histidine kinase	Bacteria	Alkaliphilus metalliredigens (strain QYMF)	1	0	**22.8**	0.36
		L7MMA2_OESDE	RIC-3	Eukaryota	Oesophagostomum dentatum	1	0	**22.4**	0.48
		C6H7B9_AJECH	Sec14 cytosolic factor	Eukaryota	Ajellomyces capsulatus (strain H143) (Darling's disease fungus) (Histoplasma capsulatum)	1	0	**22.3**	0.53
		L9VVP2_9EURY	ATPase AAA containing von Willebrand factor type A (VWA) domain-like protein	Archaea	Natronorubrum tibetense GA33	1	0	**22.2**	0.57
		B5RV46_DEBHA	DEHA2G07304p	Eukaryota	Debaryomyces hansenii (strain ATCC 36239 / CBS 767 / JCM 1990 / NBRC 0083 / IGC 2968)	1	0	**22.0**	0.62
		C5DTL9_ZYGRC	ZYRO0C09614p	Eukaryota	Zygosaccharomyces rouxii (strain ATCC 2623 / CBS 732 / NBRC 1130 / NCYC 568 / NRRL Y-229)	1	0	**21.8**	0.76
		K2HQK2_ENTNP	Major facilitator superfamily protein	Eukaryota	Entamoeba nuttalli (strain P19) (Amoeba)	1	0	**21.7**	0.77
		W4GC94_9STRA	tRNA pseudouridine(55) synthase	Eukaryota	Aphanomyces astaci	1	0	**21.5**	0.91
M-*ORF*	Pfamseq	F0U682_AJEC8	SEC14 cytosolic factor	Eukaryota	Ajellomyces capsulatus (strain H88) (Darling's disease fungus) (Histoplasma capsulatum)	1	0	**21.5**	0.94
		H0DG78_9STAP	Nuclease-like protein	Bacteria	Staphylococcus pettenkoferi VCU012	1	0	**21.4**	0.99
F-*ORF*	UniProtKB	F4ZFW9_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	8	8	**122.9**	**1.8E-32**
		F4ZFX0_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	7	7	**110.8**	**1.1E-28**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZGF0_9BIVA	H open reading frame	Eukaryota	Lasmigona subviridis	7	7	**109.6**	**2.5E-28**
		F4ZFW8_9BIVA	H-orf protein	Eukaryota	Lasmigona subviridis	7	7	**109.6**	**2.6E-28**
		F4ZFN3_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma lividus	1	1	**108.3**	**6.6E-28**
		F4ZFH5_MARMG	Female-specific orf protein	Eukaryota	Margaritifera margaritifera	1	1	**101.6**	**7.9E-26**
		F4ZFV5_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**92.4**	**5.9E-23**
		F4ZFW3_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**92.1**	**7.6E-23**
		F4ZFG1_LAMSI	Female-specific orf protein	Eukaryota	Lampsilis siliquoidea	1	1	**89.5**	**4.8E-22**
		F4ZFV6_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**89.1**	**6.6E-22**
		F4ZFG0_9BIVA	Female-specific orf protein	Eukaryota	Lampsilis powellii	1	1	**89.0**	**7.0E-22**
		F4ZFW0_LASCM	H-orf protein (Fragment)	Eukaryota	Lasmigona compressa	7	7	**86.9**	**3.0E-21**
		F4ZFW4_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**84.9**	**1.3E-20**
		F4ZFV7_9BIVA	Female-specific orf protein	Eukaryota	Villosa iris	1	1	**82.1**	**9.4E-20**
		F4ZFW7_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**79.6**	**5.7E-19**
		F4ZGB5_LASCM	H open reading frame	Eukaryota	Lasmigona compressa	5	5	**72.1**	**1.3E-16**
		F4ZFE6_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**72.0**	**1.4E-16**
		F4ZFF2_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**70.8**	**3.3E-16**
		V9PBQ9_9BIVA	F-ORF	Eukaryota	Solenaia carinatus	1	1	**70.6**	**3.9E-16**
		F4ZFH4_LEMRI	Female-specific orf protein	Eukaryota	Lemiox rimosus	1	1	**69.4**	**9.0E-16**
		F4ZFL6_9BIVA	Female-specific orf protein	Eukaryota	Quadrula houstonensis	1	1	**69.0**	**1.2E-15**
		F4ZFF9_9BIVA	Female-specific orf protein	Eukaryota	Echyridella menziesii	1	1	**68.9**	**1.2E-15**
		F4ZFF3_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**68.4**	**1.8E-15**
		F4ZFH7_9BIVA	Female-specific orf protein	Eukaryota	Margaritifera marrianae	1	1	**67.4**	**3.7E-15**
		F4ZFF4_9BIVA	Female-specific orf protein	Eukaryota	Ellipsaria lineolata	1	1	**66.5**	**7.2E-15**
F-*ORF*	UniProtKB	F4ZFK9_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**66.1**	**9.8E-15**
		F4ZFI6_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**66.0**	**1.0E-14**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFK7_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**65.5**	**1.4E-14**
		F4ZFI9_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**65.4**	**1.6E-14**
		F4ZFI8_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**65.3**	**1.6E-14**
		F4ZFL2_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**65.3**	**1.6E-14**
		F4ZFW1_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	6	4	**64.8**	**2.4E-14**
		F4ZFI1_9BIVA	Female-specific orf protein	Eukaryota	Potamilus metnecktayi	1	1	**64.7**	**2.6E-14**
		F4ZFF6_FUSFL	Female-specific orf protein	Eukaryota	Fusconaia flava	1	1	**64.6**	**2.9E-14**
		F4ZFK6_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**63.7**	**5.3E-14**
		F4ZFE3_9BIVA	Female-specific orf protein	Eukaryota	Alasmidonta undulata	1	1	**63.3**	**6.9E-14**
		F4ZFW6_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**63.2**	**7.7E-14**
		X2CT99_9BIVA	H open reading frame	Eukaryota	Dahurinaia dahurica	1	1	**62.6**	**1.2E-13**
		F4ZFL4_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**62.5**	**1.3E-13**
		F4ZFK8_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**62.4**	**1.3E-13**
		F4ZFG2_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**61.8**	**2.1E-13**
		F4ZFE8_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**61.5**	**2.5E-13**
		F4ZFG3_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**61.1**	**3.6E-13**
		F4ZFN6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**60.7**	**4.5E-13**
		F4ZFN5_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**60.5**	**5.4E-13**
		F4ZFW5_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**60.4**	**5.7E-13**
		F4ZFQ4_9BIVA	Female-specific orf protein	Eukaryota	Truncilla macrodon	1	1	**60.3**	**6.2E-13**
		F4ZFN4_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**59.1**	**1.4E-12**
		F4ZFF5_9BIVA	Female-specific orf protein	Eukaryota	Fusconaia ebenus	1	1	**58.5**	**2.2E-12**
		F4ZFQ2_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma texasiensis	1	1	**57.8**	**3.7E-12**
		F4ZFE2_ALAMA	Female-specific orf protein	Eukaryota	Alasmidonta marginata	1	1	**57.5**	**4.6E-12**
		F4ZFH0_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**57.5**	**4.6E-12**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFI3_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**57.1**	**6.2E-12**
		F4ZFP1_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**56.2**	**1.2E-11**
F-*ORF*	UniProtKB	F4ZFP9_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**56.2**	**1.2E-11**
		F4ZFW2_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	2	2	**55.9**	**1.5E-11**
		F4ZFY0_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.6**	**1.8E-11**
		F4ZFH3_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**55.6**	**1.9E-11**
		F4ZFX1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.5**	**2.0E-11**
		F4ZFY3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.4**	**2.1E-11**
		F4ZFX3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	2	2	**55.3**	**2.2E-11**
		F4ZFY2_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.2**	**2.4E-11**
		F4ZFX2_MARFC	H open reading frame	Eukaryota	Margaritifera falcata	1	1	**55.1**	**2.6E-11**
		F4ZFY1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**53.8**	**6.3E-11**
		F4ZFT3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**53.3**	**9.3E-11**
		F4ZFU5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**53.0**	**1.2E-10**
		F4ZFP6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**52.5**	**1.6E-10**
		F4ZG87_9BIVA	F-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**51.6**	**3.2E-10**
		F4ZFI0_9BIVA	Female-specific orf protein	Eukaryota	Megalonaias nervosa	1	1	**51.6**	**3.3E-10**
		F4ZFT2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**50.7**	**6.1E-10**
		F4ZFT5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**50.7**	**6.3E-10**
		F4ZFR5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**49.4**	**1.5E-09**
		F4ZFR3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.7**	**2.5E-09**
		F4ZFQ5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.5**	**2.9E-09**
		F4ZFS9_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.5**	**3.1E-09**
		F4ZFR2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.1**	**3.8E-09**
		F4ZFQ6_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**46.8**	**9.7E-09**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFQ8_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**46.8**	**1.0E-08**
		F4ZFT0_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**46.8**	**1.0E-08**
		F4ZFU7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**45.8**	**2.1E-08**
		F4ZFL9_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**45.4**	**2.7E-08**
		F4ZFM7_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma glans	1	1	**44.5**	**5.4E-08**
		F4ZFL8_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**44.0**	**7.5E-08**
F-*ORF*	UniProtKB	F4ZFM0_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**43.4**	**1.2E-07**
		F4ZFQ7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**42.3**	**2.5E-07**
		U5KJG1_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**42.0**	**3.0E-07**
		U5KJ96_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**41.2**	**5.7E-07**
		U5KJC3_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**39.3**	**2.1E-06**
		F2WZ99_SINWO	F ORF	Eukaryota	Sinanodonta woodiana	1	1	**33.0**	**0.0002**
		Q6D9Z7_PECAS	Putative membrane protein	Bacteria	Pectobacterium atrosepticum (strain SCRI 1043 / ATCC BAA-672)	2	0	**21.5**	0.76
	PDB	3tia_A	Neuraminidase	Viruses	Influenza A virus (A/RI/5+/1957(H2N2))	1	0	14.7	0.56
	QfO	Q59ZX2_CANAL	FTR1 family protein	Eukaryota	Candida albicans (strain SC5314 / ATCC MYA-2876)	3	0	18.6	0.098
		A2ERK8_TRIVA	DNA polymerase epsilon. catalytic subunit, putative	Eukaryota	Trichomonas vaginalis	2	0	16.1	0.59
		Q9VE38_DROME	CG14302	Eukaryota	Drosophila melanogaster	1	0	16.0	0.64
	Pfamseq	F4ZFW9_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	8	8	**122.9**	**2.9E-32**
		F4ZFX0_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	7	7	**110.8**	**1.8E-28**
		F4ZGF0_9BIVA	H open reading frame	Eukaryota	Lasmigona subviridis	7	7	**109.6**	**4.0E-28**
		F4ZFW8_9BIVA	H-orf protein	Eukaryota	Lasmigona subviridis	7	7	**109.6**	**4.3E-28**
		F4ZFN3_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma lividus	1	1	**108.3**	**1.1E-27**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFH5_MARMG	Female-specific orf protein	Eukaryota	Margaritifera margaritifera	1	1	**101.6**	**1.3E-25**
		F4ZFV5_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**92.4**	**9.5E-23**
		F4ZFW3_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**92.1**	**1.2E-22**
		F4ZFG1_LAMSI	Female-specific orf protein	Eukaryota	Lampsilis siliquoidea	1	1	**89.5**	**7.8E-22**
		F4ZFV6_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**89.1**	**1.1E-21**
		F4ZFG0_9BIVA	Female-specific orf protein	Eukaryota	Lampsilis powellii	1	1	**89.0**	**1.1E-21**
		F4ZFW0_LASCM	H-orf protein (Fragment)	Eukaryota	Lasmigona compressa	7	7	**86.9**	**4.8E-21**
		F4ZFW4_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**84.9**	**2.1E-20**
		F4ZFV7_9BIVA	Female-specific orf protein	Eukaryota	Villosa iris	1	1	**82.1**	**1.5E-19**
		F4ZFW7_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**79.6**	**9.1E-19**
F-*ORF*	Pfamseq	F4ZGB5_LASCM	H open reading frame	Eukaryota	Lasmigona compressa	5	5	**72.1**	**2.1E-16**
		F4ZFE6_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**72.0**	**2.3E-16**
		F4ZFF2_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**70.8**	**5.3E-16**
		V9PBQ9_9BIVA	F-ORF	Eukaryota	Solenaia carinatus	1	1	**70.6**	**6.3E-16**
		F4ZFH4_LEMRI	Female-specific orf protein	Eukaryota	Lemiox rimosus	1	1	**69.4**	**1.4E-15**
		F4ZFL6_9BIVA	Female-specific orf protein	Eukaryota	Quadrula houstonensis	1	1	**69.0**	**1.9E-15**
		F4ZFF9_9BIVA	Female-specific orf protein	Eukaryota	Echyridella menziesii	1	1	**68.9**	**2.0E-15**
		F4ZFF3_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**68.4**	**3.0E-15**
		F4ZFH7_9BIVA	Female-specific orf protein	Eukaryota	Margaritifera marrianae	1	1	**67.4**	**6.0E-15**
		F4ZFF4_9BIVA	Female-specific orf protein	Eukaryota	Ellipsaria lineolata	1	1	**66.5**	**1.2E-14**
		F4ZFK9_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**66.1**	**1.6E-14**
		F4ZFI6_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**66.0**	**1.7E-14**
		F4ZFK7_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**65.5**	**2.3E-14**
		F4ZFI9_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**65.4**	**2.6E-14**
		F4ZFI8_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**65.3**	**2.7E-14**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFL2_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**65.3**	**2.7E-14**
		F4ZFW1_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	6	4	**64.8**	**3.9E-14**
		F4ZFI1_9BIVA	Female-specific orf protein	Eukaryota	Potamilus metnecktayi	1	1	**64.7**	**4.1E-14**
		F4ZFF6_FUSFL	Female-specific orf protein	Eukaryota	Fusconaia flava	1	1	**64.6**	**4.6E-14**
		F4ZFK6_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**63.7**	**8.6E-14**
		F4ZFE3_9BIVA	Female-specific orf protein	Eukaryota	Alasmidonta undulata	1	1	**63.3**	**1.1E-13**
		F4ZFW6_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**63.2**	**1.2E-13**
		X2CT99_9BIVA	H open reading frame	Eukaryota	Dahurinaia dahurica	1	1	**62.6**	**1.9E-13**
		F4ZFL4_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**62.5**	**2.0E-13**
		F4ZFK8_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**62.4**	**2.2E-13**
		F4ZFG2_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**61.8**	**3.4E-13**
		F4ZFE8_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**61.5**	**4.1E-13**
		F4ZFG3_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**61.1**	**5.7E-13**
		F4ZFN6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**60.7**	**7.3E-13**
F-*ORF*	Pfamseq	F4ZFN5_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**60.5**	**8.6E-13**
		F4ZFW5_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**60.4**	**9.1E-13**
		F4ZFQ4_9BIVA	Female-specific orf protein	Eukaryota	Truncilla macrodon	1	1	**60.3**	**1.0E-12**
		F4ZFN4_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**59.1**	**2.3E-12**
		F4ZFF5_9BIVA	Female-specific orf protein	Eukaryota	Fusconaia ebenus	1	1	**58.5**	**3.5E-12**
		F4ZFQ2_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma texasiensis	1	1	**57.8**	**5.9E-12**
		F4ZFE2_ALAMA	Female-specific orf protein	Eukaryota	Alasmidonta marginata	1	1	**57.5**	**7.4E-12**
		F4ZFH0_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**57.5**	**7.4E-12**
		F4ZFI3_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**57.1**	**1.0E-11**
		F4ZFP1_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**56.2**	**1.9E-11**
		F4ZFP9_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**56.2**	**1.9E-11**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFW2_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	2	2	**55.9**	**2.4E-11**
		F4ZFH3_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**55.6**	**3.0E-11**
		F4ZFY0_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.6**	**3.0E-11**
		F4ZFX1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.5**	**3.1E-11**
		F4ZFY3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.4**	**3.3E-11**
		F4ZFX3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	2	2	**55.3**	**3.6E-11**
		F4ZFY2_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.2**	**3.9E-11**
		F4ZFX2_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**55.1**	**4.2E-11**
		F4ZFY1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**53.8**	**1.0E-10**
		F4ZFT3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**53.3**	**1.5E-10**
		F4ZFU5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**53.0**	**1.9E-10**
		F4ZFP6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**52.5**	**2.6E-10**
		F4ZG87_9BIVA	F-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**51.6**	**5.2E-10**
		F4ZFI0_9BIVA	Female-specific orf protein	Eukaryota	Megalonaias nervosa	1	1	**51.6**	**5.3E-10**
		F4ZFT2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**50.7**	**9.9E-10**
		F4ZFT5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**50.7**	**1.0E-09**
		F4ZFR5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**49.4**	**2.4E-09**
		F4ZFR3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.7**	**4.1E-09**
F-*ORF*	Pfamseq	F4ZFQ5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.5**	**4.7E-09**
		F4ZFS9_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.5**	**4.9E-09**
		F4ZFR2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**48.1**	**6.1E-09**
		F4ZFQ6_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**46.8**	**1.6E-08**
		F4ZFQ8_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**46.8**	**1.6E-08**
		F4ZFT0_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**46.8**	**1.6E-08**
		F4ZFU7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**45.8**	**3.3E-08**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFL9_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**45.4**	**4.3E-08**
		F4ZFM7_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma glans	1	1	**44.5**	**8.7E-08**
		F4ZFL8_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**44.0**	**1.2E-07**
		F4ZFM0_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**43.4**	**1.9E-07**
		F4ZFQ7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**42.3**	**4.0E-07**
		U5KJG1_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**42.0**	**4.9E-07**
		U5KJ96_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**41.2**	**9.2E-07**
		U5KJC3_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**39.3**	**3.4E-06**
		F2WZ99_SINWO	F ORF	Eukaryota	Sinanodonta woodiana	1	1	**33.0**	**3.2E-04**

NOTE : Proteins described only as “uncharacterized”, “putative”, or not annotated in general, have been removed since no information can be obtained. In bold are bit scores ≥20 and E-values ≤0.001. Results are ordered by profile, database, and E-value.

Supplementary Table X. Filtered hmmsearch output for the M-*ORF* and F-*ORF* HMM profiles built using custom parameters with hmmbuild.

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
M-*ORF*	UniProtKB	F4ZG80_9BIVA	M-specific morf protein	Eukaryota	Utterbackia peninsularis	1	1	**386.2**	**6.1E-112**
		V9PBU4_9BIVA	M-ORF	Eukaryota	Solenaia carinatus	1	1	**190.5**	**8.2E-53**
		A0A023I1I6_ANOAN	M-ORF	Eukaryota	Anodonta anatina	2	2	**117.7**	**8.2E-31**
		A0A023I1E9_ANOAN	M-ORF	Eukaryota	Anodonta anatina	2	2	**117.2**	**1.1E-30**
	SwissProt	OTCC_TREDE	Ornithine carbamoyltransferase, catabolic	Bacteria	Treponema denticola (strain ATCC 35405 / CIP 103919 / DSM 14222)	2	0	13.4	0.28
		CDSA_DICDI	Probable phosphatidate cytidylyltransferase	Eukaryota	Dictyostelium discoideum	1	0	12.5	0.51
	PDB	2ml9_A	Yop proteins translocation protein U	Bacteria	Yersinia pseudotuberculosis IP 32953	1	0	14.1	0.085
	QfO	C3Z4U7_BRAFL	Putative uncharacterized protein	Eukaryota	Branchiostoma floridae	1	0	14.7	0.16
		CDSA_DICDI	Probable phosphatidate cytidylyltransferase	Eukaryota	Dictyostelium discoideum	1	0	12.5	0.71
	Pfamseq	F4ZG80_9BIVA	M-specific morf protein	Eukaryota	Utterbackia peninsularis	1	1	**386.2**	**9.8E-112**
		V9PBU4_9BIVA	M-ORF	Eukaryota	Solenaia carinatus	1	1	**190.5**	**1.3E-52**
		A0A023I1I6_ANOAN	M-ORF	Eukaryota	Anodonta anatina	2	2	**117.7**	**1.3E-30**
		A0A023I1E9_ANOAN	M-ORF	Eukaryota	Anodonta anatina	2	2	**117.2**	**1.8E-30**
F-*ORF*	UniProtKB	F4ZFN3_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma lividus	1	1	**216.9**	**1.3E-60**
		F4ZFH5_MARMG	Female-specific orf protein	Eukaryota	Margaritifera margaritifera	1	1	**168.3**	**8.3E-46**
		F4ZFW9_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	8	8	**163.0**	**3.4E-44**
		F4ZFV6_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**149.0**	**6.2E-40**
		F4ZFX0_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	7	7	**148.9**	**6.7E-40**
		F4ZGF0_9BIVA	H open reading frame	Eukaryota	Lasmigona subviridis	7	7	**147.4**	**1.9E-39**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFW8_9BIVA	H-orf protein	Eukaryota	Lasmigona subviridis	7	7	**146.4**	**3.7E-39**
		F4ZFV5_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**133.7**	**2.7E-35**
F-*ORF*	UniProtKB	F4ZFW3_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**126.7**	**3.7E-33**
		F4ZFH4_LEMRI	Female-specific orf protein	Eukaryota	Lemiox rimosus	1	1	**124.0**	**2.4E-32**
		F4ZFW7_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**122.7**	**6.2E-32**
		F4ZFW4_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**120.2**	**3.5E-31**
		F4ZFI6_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**120.1**	**3.8E-31**
		F4ZFI9_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**119.5**	**5.8E-31**
		F4ZFG1_LAMSI	Female-specific orf protein	Eukaryota	Lampsilis siliquoidea	1	1	**119.2**	**7.2E-31**
		F4ZFI8_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**118.6**	**1.0E-30**
		F4ZFL2_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**118.6**	**1.0E-30**
		F4ZFG0_9BIVA	Female-specific orf protein	Eukaryota	Lampsilis powellii	1	1	**118.6**	**1.1E-30**
		V9PBQ9_9BIVA	F-ORF	Eukaryota	Solenaia carinatus	1	1	**117.2**	**3.0E-30**
		F4ZFV7_9BIVA	Female-specific orf protein	Eukaryota	Villosa iris	1	1	**115.7**	**8.3E-30**
		F4ZFW0_LASCM	H-orf protein (Fragment)	Eukaryota	Lasmigona compressa	7	7	**115.1**	**1.2E-29**
		F4ZFE6_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**113.0**	**5.5E-29**
		F4ZFG2_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**112.6**	**7.4E-29**
		F4ZFG3_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**111.9**	**1.2E-28**
		F4ZFK9_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**110.3**	**3.5E-28**
		F4ZFK7_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**109.5**	**6.4E-28**
		F4ZFE8_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**108.8**	**1.1E-27**
		F4ZFE3_9BIVA	Female-specific orf protein	Eukaryota	Alasmidonta undulata	1	1	**105.6**	**9.5E-27**
		F4ZFK6_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**105.3**	**1.2E-26**
		F4ZFK8_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**104.2**	**2.5E-26**
		F4ZFL4_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**103.7**	**3.6E-26**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZGB5_LASCM	H open reading frame	Eukaryota	Lasmigona compressa	4	4	**97.7**	**2.5E-24**
		F4ZFW6_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**97.6**	**2.6E-24**
		F4ZFT3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**96.8**	**4.7E-24**
		F4ZFE2_ALAMA	Female-specific orf protein	Eukaryota	Alasmidonta marginata	1	1	**96.5**	**5.5E-24**
		F4ZFH0_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**96.5**	**5.5E-24**
		F4ZFU5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**95.9**	**8.7E-24**
F-*ORF*	UniProtKB	F4ZFI3_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**93.8**	**3.9E-23**
		F4ZFF9_9BIVA	Female-specific orf protein	Eukaryota	Echyridella menziesii	1	1	**92.4**	**1.0E-22**
		F4ZG87_9BIVA	F-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**92.1**	**1.3E-22**
		F4ZFT5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**91.7**	**1.6E-22**
		F4ZFF2_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**91.1**	**2.5E-22**
		F4ZFL6_9BIVA	Female-specific orf protein	Eukaryota	Quadrula houstonensis	1	1	**90.9**	**2.8E-22**
		F4ZFT2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**90.0**	**5.3E-22**
		F4ZFU7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**88.7**	**1.3E-21**
		F4ZFH3_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**88.4**	**1.6E-21**
		F4ZFF4_9BIVA	Female-specific orf protein	Eukaryota	Ellipsaria lineolata	1	1	**88.0**	**2.2E-21**
		F4ZFW5_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**87.4**	**3.4E-21**
		F4ZFF3_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**87.0**	**4.2E-21**
		F4ZFR5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**86.5**	**6.3E-21**
		F4ZFR3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**86.0**	**8.7E-21**
		F4ZFS9_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**85.6**	**1.2E-20**
		F4ZFW2_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	2	2	**85.2**	**1.5E-20**
		F4ZFR2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**85.1**	**1.7E-20**
		F4ZFQ5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**84.8**	**2.0E-20**
		F4ZFF6_FUSFL	Female-specific orf protein	Eukaryota	Fusconaia flava	1	1	**83.5**	**5.2E-20**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFQ6_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**83.1**	**6.9E-20**
		F4ZFQ2_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma texasiensis	1	1	**82.8**	**8.1E-20**
		F4ZFT0_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**82.7**	**8.8E-20**
		F4ZFQ8_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**82.0**	**1.4E-19**
		F4ZFN6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**81.8**	**1.7E-19**
		F4ZFW1_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	4	**81.7**	**1.8E-19**
		F4ZFN5_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**80.6**	**3.9E-19**
		F4ZFQ4_9BIVA	Female-specific orf protein	Eukaryota	Truncilla macrodon	1	1	**78.7**	**1.4E-18**
		F4ZFN4_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**78.1**	**2.3E-18**
		F4ZFQ7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**74.3**	**3.1E-17**
F-*ORF*	UniProtKB	F4ZFI1_9BIVA	Female-specific orf protein	Eukaryota	Potamilus metnecktayi	1	1	**72.0**	**1.6E-16**
		F4ZFP1_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**70.9**	**3.3E-16**
		F4ZFF5_9BIVA	Female-specific orf protein	Eukaryota	Fusconaia ebenus	1	1	**68.5**	**1.8E-15**
		F4ZFP9_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**67.7**	**3.3E-15**
		F4ZFL9_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**65.0**	**2.1E-14**
		F4ZFP6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**64.6**	**2.8E-14**
		F4ZFH7_9BIVA	Female-specific orf protein	Eukaryota	Margaritifera marrianae	1	1	**64.2**	**3.8E-14**
		U5KJG1_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**63.6**	**5.9E-14**
		U5KJ96_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**61.3**	**2.9E-13**
		F4ZFL8_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**60.7**	**4.2E-13**
		F4ZFX2_MARFC	H open reading frame	Eukaryota	Margaritifera falcata	1	1	**60.3**	**5.7E-13**
		F4ZFY0_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**60.2**	**6.1E-13**
		F4ZFY3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**60.2**	**6.1E-13**
		F4ZFY2_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**59.6**	**9.1E-13**
		F4ZFX1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**59.5**	**9.8E-13**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFX3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	2	2	**59.4**	**1.1E-12**
		F4ZFM7_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma glans	1	1	**59.2**	**1.2E-12**
		X2CT99_9BIVA	H open reading frame	Eukaryota	Dahurinaia dahurica	1	1	**59.2**	**1.2E-12**
		F4ZFY1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**58.6**	**1.9E-12**
		F4ZFM0_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**58.4**	**2.2E-12**
		F4ZFI0_9BIVA	Female-specific orf protein	Eukaryota	Megalonaias nervosa	1	1	**57.3**	**4.8E-12**
		U5KJC3_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**56.4**	**9.0E-12**
		F2WZ99_SINWO	F ORF	Eukaryota	Sinanodonta woodiana	1	1	**39.2**	**1.5E-06**
		A0A0C5RBW4_9MOLU	Strain ATCC 49782 genome	Bacteria	Ureaplasma diversum	1	0	**21.0**	0.51
		A0A091H1Z3_BUCRH	Metalloreductase STEAP4 (Fragment)	Eukaryota	Buceros rhinoceros silvestris	1	0	**20.6**	0.68
		A0A091Q5Q8_LEPDC	Metalloreductase STEAP4 (Fragment)	Eukaryota	Leptosomus discolor	1	0	**20.6**	0.69
		C0F8Y4_9RICK	Efflux transporter, RND family, MFP subunit (Fragment)	Bacteria	Wolbachia endosymbiont of Muscidifurax uniraptor	1	0	**20.6**	0.7
F-*ORF*	SwissProt	NRAM_I68A6	Neuraminidase	Viruses	Influenza A virus (A/Northern Territories/60-JY2/1968(H3N2))	1	0	16.1	0.17
		NRAM_I57A5	Neuraminidase	Viruses	Influenza A virus (strain A/Singapore/1/1957 H2N2)	1	0	14.9	0.4
		NRAM_I60A0	Neuraminidase	Viruses	Influenza A virus (strain A/Ann Arbor/6/1960 H2N2)	1	0	14.9	0.41
		NRAM_I66A1	Neuraminidase	Viruses	Influenza A virus (strain A/Turkey/Wisconsin/1/1966 H9N2)	1	0	14.3	0.6
		NRAM_I68A5	Neuraminidase	Viruses	Influenza A virus (A/(Puerto Rico/8/1934-Korea/426/1968)(H2N2))	1	0	14.3	0.62
		NRAM_I67A0	Neuraminidase	Viruses	Influenza A virus (strain A/Tokyo/3/1967 H2N2)	1	0	14.3	0.62
	PDB	3tia_A	Neuraminidase	Viruses	Influenza A virus (A/RI/5+/1957(H2N2))	1	0	14.9	0.2

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
	QfO	P73901_SYNY3	50S ribosomal protein L12 honologue	Bacteria	Synechocystis sp. (strain PCC 6803 / Kazusa)	1	0	15.9	0.28
	Pfamseq	F4ZFN3_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma lividus	1	1	**216.9**	**2.2E-60**
		F4ZFH5_MARMG	Female-specific orf protein	Eukaryota	Margaritifera margaritifera	1	1	**168.3**	**1.3E-45**
		F4ZFW9_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	8	8	**163.0**	**5.4E-44**
		F4ZFV6_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**149.0**	**1.0E-39**
		F4ZFX0_9BIVA	H-orf protein (Fragment)	Eukaryota	Lasmigona subviridis	7	7	**148.9**	**1.1E-39**
		F4ZGF0_9BIVA	H open reading frame	Eukaryota	Lasmigona subviridis	7	7	**147.4**	**3.0E-39**
		F4ZFW8_9BIVA	H-orf protein	Eukaryota	Lasmigona subviridis	7	7	**146.4**	**6.0E-39**
		F4ZFV5_VENEL	Female-specific orf protein	Eukaryota	Venustaconcha ellipsiformis	1	1	**133.7**	**4.3E-35**
		F4ZFW3_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**126.7**	**5.9E-33**
		F4ZFH4_LEMRI	Female-specific orf protein	Eukaryota	Lemiox rimosus	1	1	**124.0**	**3.9E-32**
		F4ZFW7_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**122.7**	**9.9E-32**
		F4ZFW4_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	5	**120.2**	**5.6E-31**
		F4ZFI6_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**120.1**	**6.2E-31**
		F4ZFI9_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**119.5**	**9.3E-31**
		F4ZFG1_LAMSI	Female-specific orf protein	Eukaryota	Lampsilis siliquoidea	1	1	**119.2**	**1.2E-30**
F-*ORF*	Pfamseq	F4ZFI8_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**118.6**	**1.7E-30**
		F4ZFL2_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**118.6**	**1.7E-30**
		F4ZFG0_9BIVA	Female-specific orf protein	Eukaryota	Lampsilis powellii	1	1	**118.6**	**1.7E-30**
		V9PBQ9_9BIVA	F-ORF	Eukaryota	Solenaia carinatus	1	1	**117.2**	**4.8E-30**
		F4ZFV7_9BIVA	Female-specific orf protein	Eukaryota	Villosa iris	1	1	**115.7**	**1.3E-29**
		F4ZFW0_LASCM	H-orf protein (Fragment)	Eukaryota	Lasmigona compressa	7	7	**115.1**	**2.0E-29**
		F4ZFE6_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**113.0**	**8.8E-29**
		F4ZFG2_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**112.6**	**1.2E-28**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFG3_LASCO	Female-specific orf protein	Eukaryota	Lasmigona complanata	1	1	**111.9**	**1.9E-28**
		F4ZFK9_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**110.3**	**5.6E-28**
		F4ZFK7_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**109.5**	**1.0E-27**
		F4ZFE8_CUMMO	Female-specific orf protein	Eukaryota	Cumberlandia monodonta	1	1	**108.8**	**1.7E-27**
		F4ZFE3_9BIVA	Female-specific orf protein	Eukaryota	Alasmidonta undulata	1	1	**105.6**	**1.5E-26**
		F4ZFK6_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**105.3**	**1.9E-26**
		F4ZFK8_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**104.2**	**4.0E-26**
		F4ZFL4_9BIVA	Female-specific orf protein	Eukaryota	Pyganodon lacustris	1	1	**103.7**	**5.7E-26**
		F4ZGB5_LASCM	H open reading frame	Eukaryota	Lasmigona compressa	4	4	**97.7**	**4.0E-24**
		F4ZFW6_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**97.6**	**4.1E-24**
		F4ZFT3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**96.8**	**7.5E-24**
		F4ZFE2_ALAMA	Female-specific orf protein	Eukaryota	Alasmidonta marginata	1	1	**96.5**	**8.8E-24**
		F4ZFH0_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**96.5**	**8.8E-24**
		F4ZFU5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**95.9**	**1.4E-23**
		F4ZFI3_PYGGR	Female-specific orf protein	Eukaryota	Pyganodon grandis	1	1	**93.8**	**6.2E-23**
		F4ZFF9_9BIVA	Female-specific orf protein	Eukaryota	Echyridella menziesii	1	1	**92.4**	**1.6E-22**
		F4ZG87_9BIVA	F-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**92.1**	**2.0E-22**
		F4ZFT5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**91.7**	**2.6E-22**
		F4ZFF2_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**91.1**	**4.0E-22**
		F4ZFL6_9BIVA	Female-specific orf protein	Eukaryota	Quadrula houstonensis	1	1	**90.9**	**4.5E-22**
		F4ZFT2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**90.0**	**8.6E-22**
F-*ORF*	Pfamseq	F4ZFU7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peninsularis	1	1	**88.7**	**2.1E-21**
		F4ZFH3_9BIVA	Female-specific orf protein	Eukaryota	Lasmigona costata	1	1	**88.4**	**2.6E-21**
		F4ZFF4_9BIVA	Female-specific orf protein	Eukaryota	Ellipsaria lineolata	1	1	**88.0**	**3.5E-21**
		F4ZFW5_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	3	3	**87.4**	**5.5E-21**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
		F4ZFF3_CYCTU	Female-specific orf protein	Eukaryota	Cyclonaias tuberculata	1	1	**87.0**	**6.8E-21**
		F4ZFR5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**86.5**	**1.0E-20**
		F4ZFR3_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**86.0**	**1.4E-20**
		F4ZFS9_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**85.6**	**1.9E-20**
		F4ZFW2_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	2	2	**85.2**	**2.4E-20**
		F4ZFR2_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**85.1**	**2.7E-20**
		F4ZFQ5_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**84.8**	**3.2E-20**
		F4ZFF6_FUSFL	Female-specific orf protein	Eukaryota	Fusconaia flava	1	1	**83.5**	**8.4E-20**
		F4ZFQ6_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**83.1**	**1.1E-19**
		F4ZFQ2_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma texasiensis	1	1	**82.8**	**1.3E-19**
		F4ZFT0_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**82.7**	**1.4E-19**
		F4ZFQ8_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**82.0**	**2.3E-19**
		F4ZFN6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**81.8**	**2.7E-19**
		F4ZFW1_LASCM	H-orf protein	Eukaryota	Lasmigona compressa	5	4	**81.7**	**2.8E-19**
		F4ZFN5_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**80.6**	**6.3E-19**
		F4ZFQ4_9BIVA	Female-specific orf protein	Eukaryota	Truncilla macrodon	1	1	**78.7**	**2.3E-18**
		F4ZFN4_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**78.1**	**3.6E-18**
		F4ZFQ7_9BIVA	Female-specific orf protein	Eukaryota	Utterbackia peggyae	1	1	**74.3**	**5.0E-17**
		F4ZFI1_9BIVA	Female-specific orf protein	Eukaryota	Potamilus metnecktayi	1	1	**72.0**	**2.6E-16**
		F4ZFP1_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**70.9**	**5.4E-16**
		F4ZFF5_9BIVA	Female-specific orf protein	Eukaryota	Fusconaia ebenus	1	1	**68.5**	**2.9E-15**
		F4ZFP9_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**67.7**	**5.4E-15**
		F4ZFL9_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**65.0**	**3.4E-14**
		F4ZFP6_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma paulus	1	1	**64.6**	**4.6E-14**
		F4ZFH7_9BIVA	Female-specific orf protein	Eukaryota	Margaritifera marrianae	1	1	**64.2**	**6.1E-14**

Profile	*Database*	*Target*	*Description*	*Kingdom*	*Species*	*# hits*	*# significant hits*	*Bit Score*	*E-value*
F-*ORF*	Pfamseq	U5KJG1_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**63.6**	**9.5E-14**
		U5KJ96_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**61.3**	**4.7E-13**
		F4ZFL8_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**60.7**	**6.8E-13**
		F4ZFX2_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**60.3**	**9.2E-13**
		F4ZFY0_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**60.2**	**9.9E-13**
		F4ZFY3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**60.2**	**9.9E-13**
		F4ZFY2_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**59.6**	**1.5E-12**
		F4ZFX1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**59.5**	**1.6E-12**
		F4ZFX3_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	2	2	**59.4**	**1.8E-12**
		F4ZFM7_9BIVA	Female-specific orf protein	Eukaryota	Toxolasma glans	1	1	**59.2**	**2.0E-12**
		X2CT99_9BIVA	H open reading frame	Eukaryota	Dahurinaia dahurica	1	1	**59.2**	**2.0E-12**
		F4ZFY1_MARFC	H-orf protein	Eukaryota	Margaritifera falcata	1	1	**58.6**	**3.0E-12**
		F4ZFM0_9BIVA	Female-specific orf protein	Eukaryota	Strophitus undulatus	1	1	**58.4**	**3.5E-12**
		F4ZFI0_9BIVA	Female-specific orf protein	Eukaryota	Megalonaias nervosa	1	1	**57.3**	**7.7E-12**
		U5KJC3_ANOAN	F-ORF	Eukaryota	Anodonta anatina	1	1	**56.4**	**1.4E-11**
		F2WZ99_SINWO	F ORF	Eukaryota	Sinanodonta woodiana	1	1	**39.2**	**2.4E-06**

NOTE : parameters: `--fast --symfrac 0 --fragthresh 0 --wnone --enone`; see HMMER User's Guide at ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf for details on the commands. Proteins described only as "uncharacterized", "putative", or not annotated in general, have been removed since no information can be obtained. In bold are bit scores ≥20 and E-values ≤0.001. Results are ordered by profile, database, and E-value.

Supplementary Table XI. *Venustaconcha ellipsiformis* M-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		158-164	99.11
TIGR04294 prepilin-type processing-associated H-X9-DG domain		26-29	99.01
TIGR01167 LPXTG cell wall anchor domain		17-39	98.99
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		40-46	97.60
Cutaneous T-cell lymphoma-associated antigen 1 isoform 1	*Homo sapiens*	21-171	97.99
CTAGE family, member 5 isoform 2	*Homo sapiens*	13-171	97.92
CTAGE family, member 5 isoform 1	*Homo sapiens*	13-171	97.73
CTAGE family, member 5 isoform 4	*Homo sapiens*	21-171	97.71
TIGR03501 GlyGly-CTERM domain (rank 14)		22-35	96.47
CTAGE family, member 5 isoform 3	*Homo sapiens*	21-171	97.24
CTAGE family, member 5	*Mus musculus*	21-171	97.13
Nuclear Pore complex Protein family member (npp-11)	*Caenorhabditis elegans*	62-217	97.08
Nuclear Pore complex Protein family member (npp-11)	*Caenorhabditis elegans*	81-192	96.95
Essential subunit of the nuclear pore complex (NPC)		87-218	96.25
Subunit of the Nsp1p-Nup57p-Nup49p-Nic96p subcomplex of the nuclear pore complex (NPC)	*Saccharomyces cerevisiae*	90-192	95.37
Flagellar motor protein	*Agrobacterium tumefaciens*	11-176	95.34
Essential component of the nuclear pore complex	*Saccharomyces cerevisiae*	44-170	95.24
Collectin sub-family member 12 isoform II	*Homo sapiens*	19-171	95.15
Essential component of the nuclear pore complex	*Saccharomyces cerevisiae*	44-196	95.07
CD207 antigen, langerin	*Mus musculus*	19-149	94.92
Peroxisomal membrane protein that is a central component of the peroxisomal protein import machinery	*Saccharomyces cerevisiae*	27-167	94.76
Structural constituent of nuclear pore	*Arabidopsis thaliana*	61-217	94.64
Fc fragment of IgE, low affinity II, receptor for (CD23A)	*Homo sapiens*	24-157	94.54
Structural constituent of nuclear pore	*Arabidopsis thaliana*	44-172	94.40
Scavenger receptor class A, member 3	*Mus musculus*	19-171	94.36
Keratin 9	*Homo sapiens*	81-171	94.29
Nucleoporin 62kDa	*Homo sapiens*	90-218	93.89
Nucleoporin 62kDa	*Homo sapiens*	90-218	93.89

Nucleoporin 62kDa	*Homo sapiens*	90-218	93.89
Nucleoporin 62kDa	*Homo sapiens*	90-218	93.89
CTAGE family, member 5 isoform 4	*Homo sapiens*	21-171	93.64
TIGR00756 pentatricopeptide repeat domain		120-136	92.71
F02E8.5	*Caenorhabditis elegans*	43-190	93.56
CG4898-PF, isoform F	*Drosophila melanogaster*	91-171	93.56
CTAGE family, member 5	*Mus musculus*	22-183	93.52
CG4898-PK, isoform K	*Drosophila melanogaster*	90-171	93.50
Keratin complex 1, acidic, gene 9	*Mus musculus*	81-218	93.44
Cancer susceptibility candidate 4 isoform b	*Homo sapiens*	21-183	93.43
Macrophage galactose N-acetyl-galactosamine specific lectin 2		20-171	93.40
Laminin, beta 4	*Homo sapiens*	41-218	93.39
Laminin, beta 1 precursor	*Homo sapiens*	41-218	93.08
Vacuolar protein sorting 37C	*Mus musculus*	72-171	93.04
PaREP5a	*Pyrobaculum aerophilum*	89-171	92.83
PaREP5a	*Pyrobaculum aerophilum*	81-171	92.80
Laminin B1 subunit 1	*Mus musculus*	41-218	92.77
Keratin 3	*Homo sapiens*	81-171	92.39
Essential subunit of the nuclear pore complex (NPC)	*Saccharomyces cerevisiae*	84-218	92.38
B-cell receptor-associated protein BAP29 isoform c	*Homo sapiens*	24-157	92.13
Keratin 10	*Homo sapiens*	81-171	92.07
Nuclear Pore complex Protein family member (npp-1)	*Caenorhabditis elegans*	81-171	91.77
Nucleotide binding	*Arabidopsis thaliana*	90-213	91.70
Nuclear Pore complex Protein family member (npp-1)	*Caenorhabditis elegans*	41-201	91.62
Nuclear Pore complex Protein family member (npp-1)	*Caenorhabditis elegans*	41-220	91.50
Collectin sub-family member 12 isoform I	*Homo sapiens*	19-171	91.49
CG16932-PC, isoform C	*Drosophila melanogaster*	44-178	91.46
PaREP5a	*Pyrobaculum aerophilum*	79-171	91.21
C27D6.4c	*Caenorhabditis elegans*	81-293	91.19
Type I hair keratin KA36	*Homo sapiens*	90-178	91.00
Subunit of the Nsp1p-Nup57p-Nup49p-Nic96p subcomplex of the nuclear pore complex (NPC)	*Saccharomyces cerevisiae*	107-345	90.77
Shep3p Protein that acts as an adaptor between Myo4p and the She2p-mRNA complex	*Saccharomyces cerevisiae*	50-171	90.76
Keratin 1	*Homo sapiens*	81-177	90.64

CG16932-PC, isoform C	*Drosophila melanogaster*	44-155	90.63
APG16 autophagy 16-like isoform 1	*Homo sapiens*	75-200	90.61
Cortactin binding protein 2	*Homo sapiens*	46-183	90.59
B-cell receptor-associated protein BAP29 isoform b	*Homo sapiens*	24-155	90.58
ATP synthase subunit I	*Aeropyrum pernix K1*	54-171	90.51
CG7123-PA, isoform A	*Drosophila melanogaster*	44-214	90.39
CG7123-PB, isoform B	*Drosophila melanogaster*	44-214	90.39
Nuclear Pore complex Protein family member (npp-1)	*Caenorhabditis elegans*	45-182	90.32
Nuclear Pore complex Protein family member (npp-1)		38-171	90.20
CG8831-PA	*Drosophila melanogaster*	41-177	90.13
BLASTP			
Chromosome segregation protein SMC, common bacterial type		38-196	3.45e-05
Chromosome segregation ATPases		44-212	2.02e-04
Autophagy protein Apg6		67-180	1.77e-03
RNA polymerase Rpb1 C-terminal repeat domain-containing protein	*Blastomyces dermatitidis*	43-220	7e-05
SMC domain-containing protein	*Thermodesulfatator indicus*	72-162	0.008
LPXTG-motif cell wall anchor domain	*Bacillus cytotoxicus*	94-171	0.011
Cell wall anchor protein	*Bacillus cytotoxicus*	94-171	0.011
Viral A-type inclusion protein	*Trichomonas vaginalis*	53-111	0.031
SMC1, partial	*Brachionus calyciflorus*	52-111	0.089
Intracellular protein transport protein USO1	*Entamoeba dispar*	74-123	0.099
Chromosome segregation protein SMC	*Methanocaldococcus villosus*	65-193	0.49
PSIBLAST			
Chromosome segregation protein SMC, common bacterial type		38-196	3.45e-05
Chromosome segregation ATPases		44-212	2.02e-04
Autophagy protein Apg6		67-180	1.77e-03
Ankyrin-3	*Fukomys damarensis*	51-218	4e-04
Ankyrin-3	*Heterocephalus glaber*	51-218	5e-04
Ankyrin-3	*Pteropus alecto*	51-218	0.001
Motif Scan			
Lysine-rich region profile		63-217	12.512
I-TASSER			

Tropomyosin	*Oryctolagus cuniculus*		2.41
Smooth muscle myosin heavy chain	*Gallus gallus*		1.82
Secreted 45kDa protein	*Streptococcus pneumoniae*		1.75
General control protein GCN4 and Tropomyosin 1 α chain	*Oryctolagus cuniculus*		1.91
RhUL123	*Macacine herpesvirus 3*		0.671
Tyrosine-protein kinase Fes/Fps	*Homo sapiens*		0.597
SH3-containing GRB2-like protein 2	*Homo sapiens*		0.594
Metastasis suppressor protein 1	*Mus musculus*		0.588
Brain-specific angiogenesis inhibitor 1-associated protein 2-like protein 2	*Mus musculus*		0.588
Formin-binding protein 1	*Homo sapiens*		0.582
LEOA	*Escherichia coli*		0.575
ARF-GAP with coiled-coil, ANK repeat and PH domain-containing protein 1	*Homo sapiens*		0.574
FCH domain only protein 2	*Homo sapiens*		0.569
Brain-specific angiogenesis inhibitor 1-associated protein 2	*Homo sapiens*		0.568
Predict Protein			
Protein binding		1	
Cytoplasm			
Ankyrin-3	*Fukomys damarensis*		9e-35, 0.06
Coiled-coil domain-containing protein 6	*Homo sapiens*		1e-11
Coiled-coil domain-containing protein 6	*Mus musculus*		1e-11
Ankyrin-3 **(2)**	*Heterocephalus glaber*		2e-33, 0.24
Ankyrin-3 **(5)**	*Pteropus alecto*		2e-33-0.5
Myosin-6 **(3)**	*Mus musculus*		6e-5- 0.77
Myosin-6 **(3)**	*Rattus norvegicus*		4e-5-0.54
Myosin-7 **(3)**	*Canis familiaris*		7e-5- 0.19
Myosin-7 **(3)**	*Homo sapiens*		5e-5-0.14
Myosin-7 **(3)**	*Oryctolagus cuniculus*		5e-5-0.061
Unconventional myosin-Vc **(6)**	*Homo sapiens*		1e-11- 0.71
Myosin heavy chain, cardiac muscle isoform **(3)**	*Gallus gallus*		2e-5- 0.58
Myosin heavy chain, skeletal muscle **(2)**	*Oryctolagus cuniculus*		3e-5, 8e-4
Reticulocyte-binding protein 2 homolog a **(3)**	*Plasmodium falciparum*		8e-5- 5e-4
Atome2			

Myosin heavy chain, cardiac muscle beta isoform	*Homo sapiens*		86.43
Myosin-5A	*Gallus gallus*		76.77
Myosin heavy chain, cardiac muscle beta isoform	*Homo sapiens*		73.37
M protein	*Streptococcus pyogenes*		60.89
Protein Shroom	*Drosophila melanogaster*		59.94
Beclin-1 (Coiled Coil Domain)	*Rattus norvegicus*		59.27
Myosin-5A	*Gallus gallus*		54.59
ADP-ribosylation factor 6 (G domain, residues 13-175)	*Homo sapiens*		53.93
Rho-associated protein kinase 1 (coiled-coil domain (unp residues 535-700)) **(2)**	*Homo sapiens*		53.17, 41.06
Cell division protein ZAPB	*Escherichia coli*		51.37
Phosphoprotein	*Measles virus*		49.33
Tail needle protein gp26	*Enterobacteria phage P22*		48.96
Ras-related protein SEC4	*Saccharomyces cerevisiae*		48.96
C-JUN homodimer (leucine zipper domain, residues 272 - 315)	*Homo sapiens*		38.90
Secreted 45 kDa protein **(2)**	*Streptococcus pneumoniae*		37.17, 36.73
HP0958	*Helicobacter pylori*		41.08

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XII. *Quadrula quadrula* M-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		19-21	99.19
TIGR04294 prepilin-type processing-associated H-X9-DG domain		70-71	99.14
TIGR01167 LPXTG cell wall anchor domain		8-28	99.12
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		69-72	97.91
TIGR03501 GlyGly-CTERM domain		6-18	97.57

TIGR00756 pentatricopeptide repeat domain		48-60	93.60
CG18146-PB, isoform B	*Drosophila melanogaster*	9-36	89.69
Syndecan 3	*Mus musculus*	2-45	88.44
RCR		9-27	87.35
CG14181-PA	*Drosophila melanogaster*	13-37	83.64
CG18146-PA, isoform A	*Drosophila melanogaster*	9-82	82.12
BAS1 (PHYB activation tagged suppressor 1)	*Arabidopsis thaliana*	3-34	81.33
4_hem_cytochrm_NapC		18-38	80.73
CG13461-PA	*Drosophila melanogaster*	12-30	80.51
Sso2p: Plasma membrane t-SNARE	*Saccharomyces cerevisiae*	9-32	80.47
Sso1p: Plasma membrane t-SNARE	*Saccharomyces cerevisiae*	9-32	80.29
Syndecan 3	*Homo sapiens*	10-45	80.23
Syndecan 1 precursor	*Homo sapiens*	8-45	79.98
Syndecan 1 precursor	*Homo sapiens*	8-45	79.98
S-antigen	*Plasmodium falciparum*	1-19	79.73
Histidine kinase	*Nitrosopumilus maritimus*	1-19	76.65
Signal sequence receptor, alpha	*Homo sapiens*	11-91	74.74
RCR		9-27	74.71
COLlagen family member (col-36)	*Caenorhabditis elegans*	2-42	73.61
Signal sequence receptor, alpha	*Mus musculus*	11-91	72.27
UCP006158_SH3		9-27	72.08
Maltose:maltodextrin transport system permease	*Haloferax volcanii*	9-38	71.87
SYP124; t-SNARE	*Arabidopsis thaliana*	9-37	70.40
LCR19	*Arabidopsis thaliana*	1-21	69.65
SYP121; t-SNARE	*Arabidopsis thaliana*	9-46	68.98
LCR59	*Arabidopsis thaliana*	1-21	68.58
SYNtaxin family member (syn-2)	*Caenorhabditis elegans*	9-34	67.79
SYP121; t-SNARE	*Arabidopsis thaliana*	9-46	67.78
ZC190.8	*Caenorhabditis elegans*	3-33	67.59
Y116A8C.41	*Caenorhabditis elegans*	5-39	66.98
UCP006158_SH3		9-23	65.64
Translocation associated membrane protein		12-33	65.57
SYNtaxin family member (syn-1)	*Caenorhabditis elegans*	9-31	65.27
T01B11.3	*Caenorhabditis elegans*	9-31	63.67
Transmembrane protein	*Mycobacterium tuberculosis*		

CG16707-PB, isoform B	*Drosophila melanogaster*	10-27	63.06
CG16707-PA, isoform A	*Drosophila melanogaster*	10-27	63.06
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	2-42	63.00
CG12194-PA	*Drosophila melanogaster*	12-67	62.94
Copper ion binding / electron transporter	*Arabidopsis thaliana*	16-39	62.72
GGDEF family protein	*Beggiatoa sp. PS*	1-35	62.40
Y106G6E.2	*Caenorhabditis elegans*	10-28	62.07
Lectin, mannose-binding 2	*Mus musculus*	1-33	61.96
CG16707-PD, isoform D	*Drosophila melanogaster*	11-34	61.28
CG16707-PC, isoform C	*Drosophila melanogaster*	11-34	61.28
NHL25 (NDR1/HIN1-LIKE 25)	*Arabidopsis thaliana*	10-37	60.76
Serpentine Receptor, class X family member (srx-131)	*Caenorhabditis elegans*	10-41	60.38
C44H4.1	*Caenorhabditis elegans*	4-74	60.14
TMEM171: Transmembrane protein family 171		10-34	60.03
AC3.6	*Caenorhabditis elegans*	5-42	59.28
Opsin 1, short-wave-sensitive	*Homo sapiens*	8-41	59.09
Metal ion binding	*Arabidopsis thaliana*	9-37	58.54
ATG27: Autophagy-related protein 27		6-32	57.82
PsbI: Photosystem II reaction centre I protein		8-22	57.67
TcaA		7-23	57.22
I-TASSER			
Fumarate hydratase class II	*Mycobacterium tuberculosis*		0.560
Fumarate hydratase class II	*Mycobacterium smegmatis*		0.559
Fumarase Fum	*Mycobacterium marinum*		0.558
Adenylosuccinate lyase	*Mycobacterium smegmatis*		0.548
Argininosuccinate lyase	*Thermus thermophilus*		0.548
3-carboxy-cis,cis-muconate cycloisomerase	*Pseudomonas putida*		0.545
Virion RNA polymerase	*Bacteriophage n4*		0.542
Predict Protein			
Protein binding		42, 66, 105	
Nucleus			
Atome2			
Flavocytochrome C sulfide dehydrogenase (Flavin-binding subunit)	*Allochromatium vinosum*		62.97

NADH-quinone oxidoreductase	*Thermus thermophilus HB8*		60.23
Voltage-gated sodium channel	*Caldalkalibacillus thermarum*		42.44
NADH-quinone oxidoreductase subunit L			42.08
Alpha-catenin (dimerization and beta-catenin binding region)	*Mus musculus*		39.63
Ion transport protein	*Arcobacter butzleri*		36.63
Collagen alpha 1 fragment 84-116 of NC1	*Gallus gallus*		36.49
Apocytochrome f	*Chlamydomonas reinhardtii*		31.45
Disabled homolog 2	*Homo sapiens*		29.91
Minicollagen-5	*Hydra vulgaris*		19.36
Proline dehydrogenase	*Pseudomonas putida*		17.32

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XIII. *Pyganodon grandis* M-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		40-46	99.10
TIGR01167 LPXTG cell wall anchor domain		20-40	99.05
TIGR03304 outer membrane insertion C-terminal signal		55-59	98.89
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		144-149	97.48
TIGR03501 GlyGly-CTERM domain		28-40	97.15
TIGR00756 pentatricopeptide repeat domain		29-41	92.96
DYstrophin-like phenotype and CAPON related family member (dyc-1)	*Caenorhabditis elegans*	164-232	89.28
EGF-like-domain, multiple 5	*Mus musculus*	21-45	88.13
CG32048-PA, isoform A	*Drosophila melanogaster*	164-232	87.99
CG2086-PB, isoform B	*Drosophila melanogaster*	24-128	84.37
Phosphoserine aminotransferase, PSAT	*Bacillus circulans,*	104-140	79.56

Negative regulator of septation ring formation	*Lactobacillus casei*	25-60	78.48
Golgi phosphoprotein 4	*Mus musculus*	25-47	78.23
CG17213-PA	*Drosophila melanogaster*	2-55	76.37
Phosphoserine aminotransferase	*Beggiatoa sp. PS*	104-140	76.27
CG6124-PA	*Drosophila melanogaster*	19-51	75.80
F55C12.5c	*Caenorhabditis elegans*	4-63	74.61
Phosphoserine aminotransferase, PSAT	*Bacillus alcalophilus*	104-140	74.12
Cell division protein	*Yersinia pestis*	20-47	73.99
Phosphoserine_aminotransferase phosphoserine aminotransferase, Methanosarcina type.		91-140	73.60
D2092.1a	*Caenorhabditis elegans*	6-74	73.12
Septation ring formation regulator EzrA	*Streptococcus pneumoniae D39*	19-52	72.71
ZK353.4	*Caenorhabditis elegans*	1-44	72.53
PsbI: Photosystem II reaction centre I protein		22-41	72.29
Phosphoserine aminotransferase	*Beggiatoa sp. PS*	46-140	70.86
Photosystem II reaction center protein I, PsbI	*Thermosynechococcus vulcanus*	22-41	69.88
Cytochrome c oxidase, cbb3-type, CcoQ subunit		22-52	69.88
Photosystem II reaction center protein I, PsbI		22-41	69.59
Homoserine kinase ThrH	*Pseudomonas aeruginosa*	116-163	69.41
GtrA		6-35	69.05
CG2086-PA, isoform A	*Drosophila melanogaster*	13-111	68.86
Transmembrane protein	*Mycobacterium tuberculosis H37Rv*	19-59	68.34
Phosphoserine aminotransferase	*Mycobacterium tuberculosis H37Rv*	104-183	68.17
Podocalyxin-like precursor isoform 2	*Homo sapiens*	21-48	68.13
ZK945.3	*Caenorhabditis elegans*	149-200	67.86
Phosphoserine aminotransferase, PSAT	*Bacillus circulans,*	104-140	67.51
Phosphoserine aminotransferase		104-183	67.36
UBN_AB: Ubinuclein conserved middle domain		149-217	67.29
CG32177-PA (SD09769P)		22-45	67.07
Photosystem II reaction center I protein	*Synechococcus sp. CC9311*	22-41	66.83
Sec-independent translocase	*Agrobacterium tumefaciens*	25-57	66.80

Myc_target_1 Myc target protein 1.		1-39	64.30
Phosphoglycolate phosphatase, PGPase	*Pyrococcus horikoshii*	116-176	64.22
Sec-independent translocase	*Escherichia coli*	25-53	63.02
Essential cell division protein	*Escherichia coli*	20-46	62.46
Podocalyxin-like precursor isoform 1	*Homo sapiens*	21-48	61.69
Nedd4 family interacting protein 2	*Homo sapiens*	20-45	61.47
Phosphoserine aminotransferase	*Homo sapiens*	104-140	61.38
Phosphoserine aminotransferase (PSAT) family		104-183	60.87
F56H1.3	*Caenorhabditis elegans*	21-47	60.78
Phosphoserine aminotransferase	*Salmonella enterica*	104-140	60.66
Penumbra	*Homo sapiens*	10-59	60.60
ACyLtransferase-like family member (acl-2)	*Caenorhabditis elegans*	149-196	59.95
Claudin domain containing 1 protein isoform a	*Homo sapiens*	3-59	59.47
Claudin domain containing 1 protein isoform a	*Homo sapiens*	3-59	59.47
Claudin domain containing 1 protein isoform a	*Homo sapiens*	3-59	59.47
Claudin domain containing 1 protein isoform a	*Homo sapiens*	3-59	59.47
Sec-independent protein translocase protein TatB	*Yersinia pestis CO92*	25-57	59.28
Phosphoserine aminotransferase, PSAT	*Bacillus alcalophilus*	104-140	58.99
Golgi phosphoprotein 4	*Homo sapiens*	29-47	58.44
Membrane protein component of ABC phosphate transporter	*Pseudomonas aeruginosa*	9-46	57.97
Photosystem II reaction center protein I		22-41	57.78
Trans-golgi network protein 2	*Homo sapiens*	28-62	57.19
Transmembrane protein	*Mycobacterium tuberculosis*	6-42	55.77
Phosphoserine aminotransferase, PSAT	*Escherichia coli*	104-140	55.21
Phosphoglycolate phosphatase	*Pyrococcus horikoshii*	80-177	54.69
CG7695-PA	*Drosophila melanogaster*	20-37	54.45
C18H2.4	*Caenorhabditis elegans*	21-60	54.21
Claudin containing domain 1	*Mus musculus*	22-59	54.19
Phosphomannomutase 1	*Homo sapiens*	116-160	54.15
Myc_target_1: Myc target protein 1		1-39	54.06
Translocation associated membrane protein		173-220	54.03
MiRP K channel accessory subunit family (mps-4)	*Caenorhabditis elegans*	6-49	54.02
MCTP-related		9-75	53.90
Sensor histidine kinase	*Bartonella henselae*	9-47	53.80

Phosphoserine aminotransferase	*Mycobacterium tuberculosis*	91-140	53.25
Phosphoserine aminotransferase	*Campylobacter jejuni*	104-140	53.11
Claudin domain containing 1 protein isoform b	*Homo sapiens*	22-59	52.75
Integral membrane protein	*Streptomyces coelicolor*	26-54	52.57
Y77E11A.12a	*Caenorhabditis elegans*	9-62	52.43
Twin arginine translocase protein A	*Frankia alni ACN14a*	25-37	52.36
Motif Scan			
Lysine-rich region profile		142-233	12.082
Bipartite nuclear localization signal profile		183-199	4.000
I-TASSER			
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		1.64
Antigen MTB48, Mycobacterial protein **(2)**	*Mycobacterium smegmatis*		1.20, 1.37
LEOA	*Escherichia coli*		1.05
Serine/threonine-protein kinase mTOR	*Homo sapiens*		1.03
Hemocyanin KLH1	*Megathura crenulata*		1.02
Accumulation associated protein	*Staphylococcus epidermidis*		1.33
Flagellar hook-associated protein	*Burkholderia pseudomallei*		0.572
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		0.563
Flagellar hook-associated protein 1	*Salmonella enterica*		0.560
Phospholipase C beta	*Meleagris gallopavo*		0.546
LEOA	*Escherichia coli*		0.540
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-3	*Homo sapiens*		0.532
Multidrug resistance protein pgp-1	*Caenorhabditis elegans*		0.521
Interferon-induced guanylate-binding protein 1	*Homo sapiens*		0.520
Predict Protein			
Protein binding		1	
Cytoplasm			
Muscle M-line assembly protein unc-89	*Harpegnathos saltator*		0.19
Conjugative transposon TraM protein	*Parabacteroides sp. 20_3*		0.70
Atome2			
Methly-coenzyme M reductase I alpha subunit	*Methanopyrus kandleri*		85.07

Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit STT3	*Saccharomyces cerevisiae*		65.74
Succinylglutamate desuccinylase	*Chromobacterium violaceum*		55.83
Matrix protein 1	*Influenza A virus*		39.16
Receptor tyrosine-protein kinase erbB-2	*Homo sapiens*		38.24
Cyclic nucleotide-gated cation channel alpha-3	*Homo sapiens*		37.26
Talin-1 (F2F3 subdomain, UNP residues 206-405)	*Mus musculus*		23.21
Proteasome-associated ATPase (Coil coil domain)	*Mycobacterium tuberculosis*		21.27
Cytochrome c oxidase subunit 1	*Thermus thermophilus*		18.60
Helix-destabilizing protein	*Enterobacteria phage T7*		18.55
Ion transport protein	*Magnetococcus marinus*		18.17
Myosin light chain	*Saccharomyces cerevisiae*		16.20
Soluble cytochrome b562, Smoothened homolog	*Homo sapiens*		15.95
Formaldehyde-activating enzyme fae	*Methylobacterium extorquens*		13.89

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XIV. *Inversidens japanensis* M-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		40-44	99.25
TIGR04294 prepilin-type processing-associated H-X9-DG domain		28-31	99.21
TIGR01167 LPXTG cell wall anchor domain		103-107	98.80
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		46-64	97.73
TIGR03501 GlyGly-CTERM domain		22-35	96.74
TIGR00756 pentatricopeptide repeat domain		1-14	94.79
CG32245-PB, isoform B	*Drosophila melanogaster*	17-63	94.34

Septum formation initiator	*Beggiatoa sp. PS*	25-47	88.31
CG32245-PA, isoform A	*Drosophila melanogaster*	17-63	87.71
CG32245-PC, isoform C	*Drosophila melanogaster*	17-63	86.87
T24B1.1	*Caenorhabditis elegans*	3-48	85.10
SYP51 (SYNTAXIN OF PLANTS 51) **(2)**	*Arabidopsis thaliana*	1-42	83.61
STOmatin family member (sto-6)	*Caenorhabditis elegans*	19-59	82.81
Photosystem II reaction center protein PsbN	*Synechococcus sp. CC9311*	15-44	79.61
STOmatin family member (sto-4)	*Caenorhabditis elegans*	21-59	79.10
	Caenorhabditis elegans	20-63	79.02
MEChanosensory abnormality family member (mec-2)	*Caenorhabditis elegans*	21-63	78.75
Stomatin isoform a	*Homo sapiens*	20-63	78.31
Podocin	*Homo sapiens*	21-63	76.82
Golgi autoantigen, golgin subfamily a, 5	*Homo sapiens*	3-48	76.72
F0F1-type ATP synthase, subunit b	*Lactobacillus casei*	20-45	76.70
SYP52 (SYNTAXIN OF PLANTS 52) **(2)**	*Arabidopsis thaliana*	1-41	76.63
Stomatin (Epb7.2)-like 3	*Mus musculus*	21-59	76.25
UNCoordinated family member (unc-1)	*Caenorhabditis elegans*	19-63	74.89
STOmatin family member (sto-2)	*Caenorhabditis elegans*	21-63	74.40
CG7635-PA	*Drosophila melanogaster*	21-59	72.95
Ring finger protein 183	*Mus musculus*	24-63	71.09
CG14644-PA	*Drosophila melanogaster*	21-63	70.71
ATP synthase F0 B subunit	*Desulfitobacterium hafniense*	23-45	69.95
Y45F3A.8	*Caenorhabditis elegans*	24-37	69.72
F0F1 ATP synthase subunit B	*Escherichia coli*	23-45	69.20
Cell divison protein FtsB	*Escherichia coli*	26-53	68.45
Sec20		23-42	68.00
ATP synthase (subunit b)	*Bacillus subtilis*	22-45	67.79
CG14736-PA, isoform A	*Drosophila melanogaster*	21-59	66.71
Cell divison protein FtsB	*Yersinia pestis CO92*	26-47	65.60
Tumor endothelial marker 8 isoform 2 precursor	*Homo sapiens*	17-42	63.80
MEChanosensory abnormality family member (mec-2)	*Caenorhabditis elegans*	19-63	62.96
DNA repair protein complementing XP-A cells		59-114	62.85
Pheromone-regulated protein, induced during cell integrity signaling	*Saccharomyces cerevisiae*	15-42	62.69

ATP synthase B/B' CF(0)		23-45	62.51
ATP synthase chain b""	*Synechococcus sp. CC9311*	22-45	62.44
S-antigen		24-42	62.44
PaTched Related family member (ptr-12)	*Caenorhabditis elegans*	2-41	62.40
Stomatin-prohibitin homolog, transmembrane	*Haloferax volcanii DS2*	22-59	62.08
D2085.6	*Caenorhabditis elegans*	2-45	62.01
Endoplasmic Reticulum-Golgi Intermediate Compartment (ERGIC)		3-41	60.13
CG10737-PA, isoform A	*Drosophila melanogaster*	26-94	59.87
Translocation associated membrane protein		67-106	59.79
Golgi membrane protein, similar to mammalian CASP		23-43	59.55
DNaJ domain family member (dnj-26)	*Caenorhabditis elegans*	4-42	59.53
ATP synthase subunit B	*Corynebacterium diphtheriae*	19-45	59.33
Stomatin-like 3	*Homo sapiens*	19-59	59.04
Anthrax toxin receptor 1	*Mus musculus*	17-42	58.78
CG10737-PB, isoform B	*Drosophila melanogaster*	26-94	58.62
STOmatin family member (sto-1)	*Caenorhabditis elegans*	17-59	58.61
ATBS14A; protein transporter	*Arabidopsis thaliana*	4-41	57.72
CG10737-PC, isoform C	*Drosophila melanogaster*	26-94	57.24
CG10737-PD, isoform D	*Drosophila melanogaster*	26-94	57.24
CG31358-PA	*Drosophila melanogaster*	21-59	57.24
Nephrosis 2 homolog, podocin	*Mus musculus*	19-59	56.75
SYP61	*Arabidopsis thaliana*	1-39	56.60
Melanocortin 2 receptor accessory protein isoform alpha	*Homo sapiens*	23-37	55.13
ATP synthase subunit B	*Bartonella henselae*	23-45	54.94
ATP synthase F0, B subunit	*Streptococcus pneumoniae*	22-45	54.79
Melanocortin 2 receptor accessory protein isoform beta	*Homo sapiens*	23-37	54.11
C35D10.8	*Caenorhabditis elegans*	4-43	54.11
CG13409-PA	*Drosophila melanogaster*	14-46	53.62
ATP synthase subunit B	*Streptomyces coelicolor*	19-45	53.22
Actin binding	*Arabidopsis thaliana*	26-38	52.92
AFH1	*Arabidopsis thaliana*	25-36	52.85
BLASTP			
ATP synthase F0 subunit B	*Lachnospiraceae bacterium*	18-116	0.57

Motif Scan			
Lysine-rich region profile		49-118	10.073
I-TASSER			
Nucleotidyltransferase	*Agrobacterium fabrum*		0.668
Bacteriorhodopsin	*Halobacterium salinarum*		0.666
Halorhodopsin	*Natronomonas pharaonis*		0.665
Deltarhodopsin	*Haloterrigena thermotolerans*		0.663
Archaerhodopsin-1	*Halorubrum chaoviator*		0.662
Archaerhodopsin-2	*Halobacterium sp. AUS-2*		0.660
Cruxrhodopsin-3	*Haloarcula vallismortis*		0.657
Halorhodopsin	*Halobacterium salinarum*		0.652
Predict Protein			
Protein binding		1-4, 43, 90, 92	
Polynucleotide binding		64	
Mitochondrion			
Atome2			
Second mitochondria-derived activator of caspases	*Homo sapiens*		76.53
Guanine nucleotide exchange factor P115RHOGEF	*Homo sapiens*		71.51
Rep (DNA-bindig domain)	*Escherichia coli*		69.62
Nuclear factor of activated T-cells	*Homo sapiens*		37.72
Antifreeze peptide SS-3	*Myoxocephalus scorpius*		34.95
Nonstructural protein 5A (BVDV NS5A)	*Bovine viral diarrhea virus*		33.37
Functional anti-apoptotic factor vBCL-2 homolog	*Human herpesvirus 8*		27.14
Ion transport protein (Pore and cytoplasmic domains)	*Alkalilimnicola ehrlichii*		25.55
Thymosin alpha-1	*Homo sapiens*		24.55
Cytochrome c oxidase subunit 1	*Thermus thermophilus*		24.36
Ion transport protein	*Magnetococcus marinus*		23.40
NHE1 isoform of Na+/H+ exchanger (Transmembrane segment VII)	*Meriones unguiculatus*		23.11
Apoptosis regulator BAK	*Homo sapiens*		18.22

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of

10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XV. *Utterbackia peninsularis* M-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		53-60	99.16
TIGR04294 prepilin-type processing-associated H-X9-DG domain		41-44	99.06
TIGR01167 LPXTG cell wall anchor domain		18-38	98.89
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		107-121	97.32
TIGR03501 GlyGly-CTERM domain		26-38	97.24
TIGR00756 pentatricopeptide repeat domain		32-39	92.61
F58B4.4	*Caenorhabditis elegans*	59-97	84.64
EGF-like-domain, multiple 5	*Mus musculus*	19-43	78.12
F13D12.5	*Caenorhabditis elegans*	15-46	77.77
T21B10.4	*Caenorhabditis elegans*	15-46	75.99
Catenin alpha-1; four helix bundle, cell adhesion	*Mus musculus*	145-158	75.96
ZK945.3	*Caenorhabditis elegans*	132-183	72.81
Cytochrome c oxidase		20-50	72.15
Transcriptional regulator	*Lactobacillus casei*	134-163	71.49
Pheromone-regulated protein, DUP240 gene family	*Saccharomyces cerevisiae*	6-85	71.19
Peptidase A24A prepilin type IV	*Candidatus Korarchaeum cryptofilum OPF8*	10-63	69.06
Mitochondrial ribosomal protein S23		44-51	65.64
CG18146-PB, isoform B	*Drosophila melanogaster*	21-47	65.59
Catenin alpha-1; four helix bundle	*Mus musculus*	145-158	61.01
Beta-lactamase	*Pseudomonas fluorescens*	143-163	60.65
Phosphatidylserine decarboxylase	*Methanosarcina mazei Go1*	19-102	60.03
Septation ring formation regulator EzrA	*Streptococcus pneumoniae*	16-42	59.30
Reticulon 1 isoform A	*Homo sapiens*	4-56	59.03
Trigger factor ribosome-binding domain (102735) SCOP seed sequence: d1w26a2		141-167	58.82
Hup-type Ni,Fe-hydrogenase cytochrome b subunit	*Desulfitobacterium*	4-78	57.93

	hafniense Y51		
Transcriptional regulator, TetR family	*Staphylococcus aureus subsp. aureus COL*	140-163	57.48
CG32048-PA, isoform A	*Drosophila melanogaster*	147-216	57.23
Essential cell division protein	*Escherichia coli K12*	18-44	56.08
Trigger factor ribosome-binding domain (102735) SCOP seed sequence: d1t11a2		141-167	56.00
F56H1.3	*Caenorhabditis elegans*	19-45	55.38
YajQ-like (89963) SCOP seed sequence: d1in0a1		143-164	54.62
Transmembrane protein	*Mycobacterium tuberculosis*	4-40	54.42
POTASSIUM VOLTAGE-GATED CHANNEL SUBFAMILY E MEMBER 1, 3.		20-38	53.73
Multi-sensor hybrid histidine kinase	*Nostoc punctiforme*	8-62	53.52
Acyltransferase	*Streptomyces coelicolor*	126-160	53.51
PELOTA_1 PELOTA RNA binding domain.		131-172	53.49
F55C12.5c	*Caenorhabditis elegans*	2-61	51.82
Y77E11A.12a	*Caenorhabditis elegans*	4-60	50.58
Transcriptional regulator MvaT, P16 subunit	*Pseudomonas aeruginosa*	130-213	49.58
Sensor histidine kinase	*Bartonella henselae*	1-50	49.57
Zinc finger, DHHC domain containing 5	*Mus musculus*	2-45	49.44
D2092.1a	*Caenorhabditis elegans*	4-84	49.18
Olfactory receptor, family 6, subfamily C, member 6	*Homo sapiens*	1-44	48.98
C18H2.4	*Caenorhabditis elegans*	20-45	48.86
Secretory carrier-associated membrane protein (SCAMP)		26-53	48.67
Thiamin diphosphate-binding fold (THDP-binding) (52518) SCOP seed sequence: d1b0pa2		91-152	48.59
PELOTA RNA binding domain		141-172	48.40
AMPC beta-Lactamase, class C	*Citrobacter freundii*	140-163	47.16
Transcriptional regulator, TetR family protein	*Streptococcus pneumoniae*	144-163	46.07
Chromatin regulatory protein SIR2		97-119	45.86
Toll-like receptor 7	*Mus musculus*	20-80	45.51
K+_transpter_TRK		20-48	45.40
Sec-independent protein translocase protein TatB	*Yersinia pestis*	23-55	44.37
Zinc finger DHHC domain-containing protein		10-51	44.23

RNA recognition motif in regulators of calcineurin (RCANs) and similar proteins.		137-159	44.00
Maltose: maltodextrin transport system permease	*Haloferax volcanii*	20-53	43.91
Phospholipid/glycerol acyltransferase	*Beggiatoa sp. PS*	131-176	43.89
Podocalyxin-like precursor isoform 2	*Homo sapiens*	19-46	43.69
Alpha amylase catalytic domain		129-164	43.63
Beta-lactamase ACT-1	*Klebsiella pneumoniae*	140-163	43.59
Metal ion binding	*Arabidopsis thaliana*	2-43	43.14
Neopullulanase, central domain	*Bacillus stearothermophilus*	129-164	42.86
Neopullulanase, central domain	*Bacillus stearothermophilus*	129-164	42.86
GH36 glycosyl hydrolase family 36 (GH36)		143-164	42.83
Trigger factor, TF; chaperone	*Thermotoga maritima*	140-167	42.76
T22E7.2	*Caenorhabditis elegans*	10-45	42.75
Zinc finger, DHHC domain containing 5	*Homo sapiens*	2-45	42.71
UCP016495		149-189	42.37
Organic solute transporter-related		3-36	41.59
Trigger_N: Bacterial trigger factor protein (TF)		140-167	41.54
SUPFAM template c.1.8 (Trans) glycosidases (51445) SCOP seed sequence: d1gcya2		141-164	41.40
Motif Scan			
Lysine-rich region profile		115-206	9.500
Bipartite nuclear localization signal profile		166-180	4.000
I-TASSER			
4-Hydroxybutyrate CoA-transferase	*Clostridium aminobutyricum*		0.605
4-hydroxybutyrate coenzyme A transferase	*Shewanella oneidensis*		0.604
4-hydroxybutyrate CoA-transferase	*Porphyromonas gingivalis*		0.603
Coenzyme A transferase	*Yersinia pestis*		0.598
4-hydroxybutyrate CoA-transferase	*Porphyromonas gingivalis*		0.565
Succinyl-CoA:acetate coenzyme A transferase	*Acetobacter aceti*		0.565
Acetyl-CoA hydrolase/transferase family protein	*Porphyromonas gingivalis*		0.553
Predict Protein			
Protein binding		41, 77, 92, 95-98, 118, 127	
Cytoplasm			

Atome2			
Vinculin	*Gallus gallus*		50.72
ADP-ribosylation factor binding protein GGA	*Homo sapiens*		42.58
80 kDa MCM3-associated protein	*Homo sapiens*		41.19
LCoR protein	*Homo sapiens*		36.28
Transcriptional repressor COPG	*Streptococcus agalactiae*		34.49
HIG1 domain family member 1B	*Homo sapiens*		29.87
Uncharacterized protein 56B (transcription repressor)	*Sulfolobus islandicus rod-shaped virus 1*		27.96
Talin-1 (F2F3 subdomain, UNP residues 206-405)	*Mus musculus*		26.08
ScpA	*Geobacillus stearothermophilus*		22.94
HIG1 domain family member 1A	*Homo sapiens*		21.45
PlnE	*Lactobacillus plantarum*		15.35

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XVI. *Solenaia carinatus* M-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		27-30	99.15
TIGR01167 LPXTG cell wall anchor domain		80-85	98.88
TIGR03304 outer membrane insertion C-terminal signal		6-7	98.75
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		60-67	97.95
TIGR03501 GlyGly-CTERM domain		21-36	97.04
TIGR00756 pentatricopeptide repeat domain		3-13	94.60
W06F12.2a	*Caenorhabditis elegans*	20-59	90.96
SAP		38-58	90.80
S-antigen		23-41	90.41
ZK973.11	*Caenorhabditis elegans*	11-43	80.48

Thioredoxin domain containing 10	*Homo sapiens*	11-43	79.74
UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase 5	*Homo sapiens*	12-51	77.75
V-type ATPase 116 kDa subunit	*Nitrosopumilus maritimus*	4-60	70.53
Thioredoxin domain containing 10	*Mus musculus*	11-113	70.01
Golgi autoantigen, golgin subfamily a, 5	*Homo sapiens*	22-47	69.90
Photosystem II reaction center protein PsbN	*Synechococcus sp. CC9311*	14-45	68.30
CG10207-PA	*Drosophila melanogaster*	23-131	67.76
PRP38 pre-mRNA processing factor 38 (yeast) domain containing B	*Homo sapiens*	4-58	64.40
CG14084-PB, isoform B	*Drosophila melanogaster*	5-40	63.61
CG14084-PA, isoform A	*Drosophila melanogaster*	5-40	63.61
ZK757.4b	*Caenorhabditis elegans*	23-45	61.92
F12B6.2b	*Caenorhabditis elegans*	22-124	59.84
Binding	*Arabidopsis thaliana*	3-58	59.77
V-type ATP synthase subunit I	*Methanopyrus kandleri*	4-60	59.58
F12B6.2a	*Caenorhabditis elegans*	10-141	59.54
Golgi membrane protein with similarity to mammalian CASP	*Saccharomyces cerevisiae*	22-41	59.34
Hydrolase, hydrolyzing O-glycosyl compounds	*Arabidopsis thaliana*	23-44	59.22
CG1622-PA	*Drosophila melanogaster*	4-44	58.83
CG17287-PA	*Drosophila melanogaster*	23-45	58.07
Sec20 is a membrane glycoprotein associated with secretory pathway		3-41	57.96
CG14181-PA	*Drosophila melanogaster*	1-56	57.83
DNA-binding transcription factor required for the activation of the GAL genes in response to galactose; repressed by Gal80p and activated by Gal3p	*Saccharomyces cerevisiae*	1-41	57.24
Y15E3A.4	*Caenorhabditis elegans*	22-89	56.55
AC3.10	*Caenorhabditis elegans*	22-45	56.54
ZK757.4a	*Caenorhabditis elegans*	23-46	56.20
Y116A8C.41	*Caenorhabditis elegans*	23-58	55.43
Actin binding	*Arabidopsis thaliana*	25-33	55.06
CG30272-PA	*Drosophila melanogaster*	22-119	54.95
Similar to S. cerevisiae PKR1	*Schizosaccharomyces pombe*	7-61	54.74

PRP38 pre-mRNA processing factor 38 (yeast) domain containing B	*Mus musculus*	4-44	54.67
Protein transporter	*Arabidopsis thaliana*	2-41	54.37
VAC_I2L		4-42	54.08
Y47H9C.2	*Caenorhabditis elegans*	14-46	53.55
Metal ion binding	*Arabidopsis thaliana*	22-46	53.52
Metal ion binding	*Arabidopsis thaliana*	22-46	53.13
Photosystem II reaction center protein N	*Nostoc punctiforme*	21-45	53.05
ATBS14A; protein transporter	*Arabidopsis thaliana*	2-41	52.99
BCL2/adenovirus E1B 19kD interacting protein 1 isoform BNIP1	*Homo sapiens*	2-52	52.89
Sarcolycans		21-42	52.86
Binding	*Arabidopsis thaliana*	4-58	52.79
T12G3.7	*Caenorhabditis elegans*	26-88	52.77
Membrane associated histidine-rich protein, MAHRP-1		1-41	52.62
Integral membrane protein	*Streptomyces coelicolor*	22-49	52.09
CG6627-PA	*Drosophila melanogaster*	21-46	51.18
Zinc finger, DHHC domain containing 15	*Mus musculus*	22-45	51.04
CG32245-PB, isoform B	*Drosophila melanogaster*	16-43	50.50
CHL00020 psbN photosystem II protein N		21-45	50.47
ZC190.8	*Caenorhabditis elegans*	21-41	50.24
PsbN: Photosystem II reaction centre N protein (psbN)		21-45	49.96
Y51F10.4b	*Caenorhabditis elegans*	5-43	49.92
ATP synthase subunit I	*Sulfolobus solfataricus*	4-60	48.84
Erf4: Golgin subfamily A member 7/ERF4 family		5-58	48.74
Photosystem I subunit III	*Synechocystis sp.*	13-42	48.61
PsbN Photosystem II reaction centre N protein (psbN)		21-45	47.86
CG8421-PB, isoform B	*Drosophila melanogaster*	20-46	47.40
Rab5ip Rab5-interacting protein (Rab5ip)		23-61	47.18
Subunit III of photosystem I reaction centre, PsaF	*Synechococcus elongatus*	13-42	47.18
Stomatin-like 3	*Homo sapiens*	18-43	47.17
BLASTP			
Transmembrane protein 72, partial	*Anas platyrhynchos*	47-129	0.95
Motif Scan			
Lysine-rich region profile		48-119	11.508
Predict Protein			

Protein binding		17-19, 43-45, 66	
Polynucleotide binding		49, 63	
Mitochondrion			
Atome2			
RAD50 ABC-ATPase (N-terminal domain)	*Pyrococcus furiosus*		84.72
Cytoplasmic FMR1-interacting protein 1	*Homo sapiens*		65.70
Chromosomal replication initiator protein dnaA	*Mycoplasma genitalium*		62.22
KIAA0380 (RGS-like domain (residues 281-490))	*Homo sapiens*		59.65
Autophagy protein 1 (coiled-coil domain)	*Saccharomyces cerevisiae*		52.37
V-type ATP synthase subunit E	*Methanocaldococcus jannaschii*		51.34
Nuclear pore complex protein Nup54 (UNP residues 346-407)	*Rattus norvegicus*		51.32
B-cell receptor-associated protein 31	*Homo sapiens*		46.96
Protein NRD1 (CTD-interacting domain, unp residues 6-151)	*Saccharomyces cerevisiae*		45.58
Photosystem I P700 chlorophyll a apoprotein A1	*Synechococcus elongatus*		42.88
Transcription factor ATF-4	*Homo sapiens*		38.86
Cytochrome c oxidase subunit 1	*Thermus thermophilus*		27.52
Bcl-2-like protein 2 (UNP residues 2-171)	*Bos taurus*		24.75
Lmo2059 protein (KVLM pore module, truncated C-terminus (UNP residues 98-233))	*Listeria monocytogenes*		24.64
Nonstructural protein 5A (BVDV NS5A)	*Bovine viral diarrhea virus*		22.35
DNA-(apurinic or apyrimidinic site) lyase	*Homo sapiens*		4.84

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XVII. *Cumberlandia monodonta* M-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		30-34	99.52
TIGR03304 outer membrane insertion C-terminal signal		51-56	99.20
TIGR01167 LPXTG cell wall anchor domain		14-35	98.77
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		13-16	97.81
TIGR03501 GlyGly-CTERM domain		16-26	97.14
TIGR00756 pentatricopeptide repeat domain		39-51	94.35
CG9552-PA	*Drosophila melanogaster*	18-60	87.49
YMF19 Plant ATP synthase F0		17-43	84.86
Secretory carrier-associated membrane protein (SCAMP)		18-38	84.05
Protein binding	*Arabidopsis thaliana*	16-54	81.54
F01E11.3	*Caenorhabditis elegans*	16-24	80.16
YMF19: Plant ATP synthase F0		17-46	69.49
ATP synthase subunit B	*Agrobacterium tumefaciens*	17-50	67.80
Selenoprotein_S: Selenoprotein S (SelS)		8-42	66.63
VP35_FiloV		13-21	63.95
CYTOCHROME C1		21-48	63.67
F56F4.7	*Caenorhabditis elegans*	57-95	62.52
Carrier	*Arabidopsis thaliana*	18-38	62.02
Syndecan		16-41	59.22
CG18146-PB, isoform B	*Drosophila melanogaster*	20-51	55.88
CG30389-PA, isoform A	*Drosophila melanogaster*	18-37	55.74
CG30389-PC, isoform C	*Drosophila melanogaster*	18-37	55.74
CG2023-PA	*Drosophila melanogaster*	17-38	55.68
CG30415-PB, isoform B	*Drosophila melanogaster*	9-39	54.50
CG30415-PA, isoform A	*Drosophila melanogaster*	9-39	54.50
SC3 (secretory carrier 3)	*Arabidopsis thaliana*	18-49	53.74
Y46G5A.26b	*Caenorhabditis elegans*	18-26	53.07
CD4.1	*Caenorhabditis elegans*	9-46	52.82
F0F1 ATP synthase subunit B'		17-50	50.26
SCAMP homolog family member (scm-1)	*Caenorhabditis elegans*	18-38	50.18
Protein binding	*Arabidopsis thaliana*	16-48	48.42

Agal		11-18	48.41
CG16707-PD, isoform D	*Drosophila melanogaster*	20-42	47.94
CG16707-PC, isoform C	*Drosophila melanogaster*	20-42	47.94
Sperm-associated cation channel 2	*Mus musculus*	15-33	47.25
BCL2/adenovirus E1B 19kD interacting protein 1 isoform BNIP1-c	*Homo sapiens*	17-38	46.86
Transmembrane protein 57	*Mus musculus*	18-37	46.30
Syntaxin 7	*Homo sapiens*	17-39	44.74
Carrier	*Arabidopsis thaliana*	18-34	44.59
Syntaxin 7	*Mus musculus*	17-39	44.47
Carrier	*Arabidopsis thaliana*	18-34	43.51
S-antigen	*Plasmodium falciparum*	18-32	42.94
Y47D7A.13	*Caenorhabditis elegans*	7-26	42.68
G-protein-linked Acetylcholine Receptor family member (gar-1)	*Caenorhabditis elegans*	18-49	41.32
Transmembrane protein 57	*Homo sapiens*	18-37	41.20
ATP synthase subunit B	*Streptomyces coelicolor*	17-50	40.92
ATP synthase subunit B	*Bartonella henselae*	17-49	40.84
Homeodomain-like (46689) SCOP seed sequence: d1hlva2		10-21	40.78
Related to Secretory carrier-associated membrane protein 2		18-34	40.31
C15A7.2	*Caenorhabditis elegans*	20-59	40.10
Zinc beta-ribbon (57783) SCOP seed sequence: d1yuaa1		25-31	39.90
Carrier	*Arabidopsis thaliana*	18-34	39.62
Secretory carrier membrane protein 1 isoform 1	*Homo sapiens*	18-34	38.98
PeRoxireDoXin family member (prdx-6)	*Caenorhabditis elegans*	21-45	38.74
E set domains (81296) SCOP seed sequence: d1eh9a1		54-64	38.39
High affinity copper uptake protein 1; HCTR1 TMDS, oligomerization, metal transport	*Homo sapiens*	15-24	37.97
F11G11.10	*Caenorhabditis elegans*	14-24	37.43
CG32177-PA	*Drosophila melanogaster*	14-42	37.06
NK inhibitory receptor precursor	*Homo sapiens*	18-41	36.95
DUF4381 Domain of unknown function (DUF4381)		16-46	36.93
CG3268-PA	*Drosophila melanogaster*	18-40	36.73

UCP014405		33-54	36.16
CG9195-PA, isoform A	*Drosophila melanogaster*	18-34	35.18
CG15673-PA	*Drosophila melanogaster*	17-41	35.13
Metal ion binding	*Arabidopsis thaliana*	41-93	34.53
Metal ion binding	*Arabidopsis thaliana*	41-93	34.53
ATP synthase subunit B	*Bartonella henselae*	17-35	34.36
SRF-like (55455) SCOP seed sequence: d1mnma_		36-50	34.35
Secretory carrier membrane protein 2	*Mus musculus*	18-34	34.20
BCL2/adenovirus E1B 19kD interacting protein 1 isoform BNIP1-b	*Homo sapiens*	17-38	33.55
BLASTP			
Plant ATP synthase F0		17-46	3.97e-03
Motif Scan			
Lysine-rich region profile		38-93	10.360
Prokaryotic membrane lipoprotein lipid attachment site profile		1-29	6.000
I-TASSER			
Enoyl-CoA hydratase EchA1	*Mycobacterium marinum*		1.76
40S ribosomal protein S4, X isoform	*Homo sapiens*		1.22
Enoyl-coA hydratase/isomerase	*Mycobacterium abscessus*		1.23
Enoyl-CoA hydratase EchA17	*Mycobacterium marinum*		1.20
Carnitinyl-CoA dehydratas	*Mycobacterium avium*		1.19
Enoyl-CoA hydratase, EchA12_1	*Mycobacterium marinum*		1.19
Enoyl-CoA hydratase/carnithine racemase	*Magnetospirillum magneticum*		1.50
Enoyl-CoA hydratase/isomerase family protein	*Bacillus anthracis*		0.882
Enoyl-CoA hydratase	*Mycobacterium smegmatis*		0.876
Methylglutaconyl-CoA hydratase	*Homo sapiens*		0.875
Enoyl-CoA hydratase echA8	*Mycobacterium tuberculosis*		0.874
Enoyl-CoA hydratase	*Mycobacterium smegmatis*		0.874
Naphthoate synthase	*Staphylococcus aureus*		0.873
2,3-dehydroadipyl-CoA hydratase	*Escherichia coli*		0.872
Predict Protein			
Protein binding		1, 18, 29, 33,	

		36, 46	
Polynucleotide binding		92	
Nucleus			
Atome2			
SVP1-like protein 2	*Kluyveromyces lactis*		78.90
Rhomboid Intramembrane Protease	*Pseudomonas aeruginosa*		63.14
U3 small nucleolar RNA-associated protein 22	*Saccharomyces cerevisiae*		43.16
DNA-directed RNA polymerase subunit alpha	*Escherichia coli*		42.29
Myosin-X (MyTH4-FERM tandem)	*Homo sapiens*		33.40
Multidrug transporter emrE	*Escherichia coli*		28.36
Actin, alpha skeletal muscle	*Oryctolagus cuniculus*		24.15
Calcium-activated potassium channel RSK2	*Rattus norvegicus*		19.69
Calmodulin	*Rattus norvegicus*		19.51
Ion transport protein	*Magnetococcus marinus*		16.41
Calmodulin	*Homo sapiens*		15.33
High affinity copper uptake protein 1	*Homo sapiens*		8.26
14-3-3 protein beta/alpha	*Mus musculus*		7.34

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XVIII. *Hyridella menziesii* M-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		107-111	99.58
TIGR03304 outer membrane insertion C-terminal signal		53-56	99.19
TIGR01167 LPXTG cell wall anchor domain		93-109	98.67
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		123-137	97.91
Apolipoprotein Apolipoprotein A1/A4/E domain		170-296	97.84
Apolipoprotein Apolipoprotein A1/A4/E domain		169-296	97.83

Apolipoprotein A-I; lipid transport; NMR	*Mus musculus*	169-296	97.66
TIGR03501 GlyGly-CTERM domain (rank 8)		98-109	96.67
a.24.1 Apolipoprotein (47162) SCOP seed sequence: d1bz4a_		182-286	97.42
Apolipoprotein A-I; lipid transport	*Mus musculus*	169-289	97.29
Apolipoprotein E **(2)**	*Homo sapiens*	167-288	97.27
Apolipoprotien		169-296	97.19
Apolipoprotein (47162) SCOP seed sequence: d1bz4a_.		170-263	97.18
Apolipoprotein A-I preproprotein	*Homo sapiens*	170-296	97.12
Apolipoprotein A-I	*Homo sapiens*	169-284	97.09
Apolipoprotein A-I preproprotein	*Homo sapiens*	169-296	97.08
Apolipoprotein A-I; four-helix bundle, lipid transport; **(3)**	*Homo sapiens*	169-296	96.92 - 97.08
Apolipoprotein A-I; four-helix bundle, lipid transport;	*Homo sapiens*	167-296	97.02
Apolipoprotein	*Homo sapiens*	170-296	96.88
Apolipoprotein E **(2)**	*Homo sapiens*	171-264	96.85
Apolipoprotein E3	*Homo sapiens*	181-295	96.79
Apolipoprotein E3	*Homo sapiens*	167-261	96.78
Y51F10.4a	*Caenorhabditis elegans*	20-296	96.48
Apolipoprotein E, APO-E	*Homo sapiens*	170-296	96.44
Y51F10.4a	*Caenorhabditis elegans*	20-289	96.35
Y51F10.4b	*Caenorhabditis elegans*	17-298	95.84
Apolipoprotein E, APO-E	*Homo sapiens*	169-264	95.77
abortive infection protein family	*Staphylococcus aureus*	11-313	95.21
TIGR00756 pentatricopeptide repeat domain		71-90	93.94
Apolipoprotein: Apolipoprotein A1/A4/E domain		167-296	94.83
Apolipoprotein: Apolipoprotein A1/A4/E domain		167-295	94.49
CG3576-PA, isoform A	*Drosophila melanogaster*	21-171	93.82
CG3576-PB, isoform B	*Drosophila melanogaster*	21-171	93.82
Apolipoprotein E, APOE4	*Homo sapiens*	169-253	93.35
apolipoprotein A-V	*Mus musculus*	168-296	93.30
Homolog of Yeast Longevity gene family member (hyl-1)	*Caenorhabditis elegans*	21-109	93.29
Homolog of Yeast Longevity gene family member (hyl-2)	*Caenorhabditis elegans*	21-174	93.24
apolipoprotein AV	*Homo sapiens*	168-263	93.12
W06F12.2a	*Caenorhabditis elegans*	20-109	92.95
Apolipoprotein E, APOE4	*Homo sapiens*	171-264	92.62

CG30394-PB, isoform B	*Drosophila melanogaster*	20-306	92.56
CG30394-PA, isoform A	*Drosophila melanogaster*	20-306	92.56
Longevity assurance homolog 4	*Mus musculus*	21-109	92.48
Apolipoprotein AV	*Homo sapiens*	169-295	92.06
Apolipoprotein A-V	*Mus musculus*	173-296	91.63
Autosomal Highly Conserved Protein	*Homo sapiens*	10-106	91.54
Longevity assurance factor 1 (LAG1)		7-109	91.38
CG30394-PB, isoform B	*Drosophila melanogaster*	20-128	90.82
CG30394-PA, isoform A	*Drosophila melanogaster*	20-128	90.82
LAG1 longevity assurance homolog 4	*Homo sapiens*	21-109	90.77
Proline-rich transmembrane protein 2	*Homo sapiens*	17-82	90.76
Translocation protein 1	*Mus musculus*	50-275	90.69
Integral membrane protein	*Streptomyces coelicolor*	20-110	90.69
Late embryogenesis abundant (plants) LEA-related		171-298	90.44
Apolipoprotein A-IV precursor	*Homo sapiens*	169-296	90.32
BLASTP/PSIBLAST			
Voltage-dependent anion channel		31-117	1.05e-04
Histone H1-like protein Hc1		213-284	3.78e-03
Microtubule-binding protein MIP-T3		166-313	8.12e-06
Periplasmic protein TonB links inner & outer membranes		201-300	7.59e-04
Cell division protein FtsN		155-286	1.31e-03
fam-a protein	*Plasmodium chabaudi chabaudi*	170-294	1e-06
Cyclin related protein	*Plasmodium chabaudi chabaudi*	168-294	4.00e-05
fam-a protein	*Plasmodium chabaudi chabaudi*	168-294	7e-05
Choline-binding protein A **(2)**	*Streptococcus pneumoniae*	179-238	7e-05
Choline-binding protein A **(2)**	*Streptococcus pneumoniae*	180-239	7e-05
Surface protein PspC **(2)**	*Streptococcus pneumoniae*	179-238	1e-04
Surface protein PspC	*Streptococcus pneumoniae*	175-309	1e-04
LPXTG-motif cell wall anchor domain protein	*Streptococcus pneumoniae*	179-238	2e-04
Choline-binding protein A	*Streptococcus pneumoniae*	182-241	2e-04
Surface protein PspC	*Streptococcus pneumoniae*	180-239	9e-04
Peptidase	*Streptococcus pneumoniae*	182-241	0.004
BLASTP			

Surface protein PspC	*Streptococcus pneumoniae*	179-238	0.007
igA FC receptor	*Streptococcus pneumoniae*	182-241	0.012
Surface protein PspC	*Streptococcus pneumoniae*	180-302	0.12
Cyclin related protein	*Plasmodium chabaudi chabaudi*	191-250	0.91
Motif Scan			
Lysine-rich region profile		168-294	24.274
Bipartite nuclear localization signal profile		170-185, 236-251, 269-284	4.000
I-TASSER			
DNA polymerase subunit gamma-1	*Homo sapiens*		1.15
Survival motor neuron protein	*Homo sapiens*		1.79, 2.63
SHERP	*Leishmania major*		1.41, 1.40
Septation ring formation regulator EZRA	*Bacillus subtilis*		1.30
Vascular apoptosis-inducing protein 1	*Crotalus atrox*		1.01
DNA (cytosine-5)-methyltransferase 1	*Zea mays*		1.15
Accumulation associated protein	*Staphylococcus epidermidis*		1.54
Tropomyosin	*Oryctolagus cuniculus*		1.28
Septation ring formation regulator EZRA	*Bacillus subtilis*		0.943
Predict Protein			
Protein binding		1-4	
Cytoplasm			
Pneumococcal surface protein C **(2)**	*Streptococcus pneumoniae*		6e-32, 2e-08
IgA-binding beta antigen **(4)**	*Streptococcus pneumoniae*		2e-30- 4e-08
Surface protein PcpC **(8)**	*Streptococcus pneumoniae*		4e-34- 0.22
Surface protein PspC **(10)**	*Streptococcus pneumoniae*		4e-34- 0.22
Titin **(15)**	*Mus musculus*		2e-09- 0.037
Muscle M-line assembly protein unc-89 **(16)**	*Caenorhabditis elegans*		1e-12- 0.018

Atome2			
Nucleoprotein (N-terminal domain (residues 1-74))	*Andes virus*		63.91
Nuclear distribution protein NUDE-like 1	*Homo sapiens*		61.83
Alpha-synuclein	*Homo sapiens*		60.92
Beclin-1	*Rattus norvegicus*		59.76
Bud site selection protein 6	*Saccharomyces cerevisiae*		56.98
Spindle and kinetochore-associated protein 3	*Homo sapiens*		48.52
Actin-related protein 7	*Saccharomyces cerevisiae*		38.22
Synaptobrevin 2	*Rattus norvegicus*		33.50
Talin-1 (Vbs2b domain, residues 787-91)	*Mus musculus*		11.89

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XIX. *Anodonta anatina* M-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		57-64	99.24
TIGR04294 prepilin-type processing-associated H-X9-DG domain		29-32	99.04
TIGR01167 LPXTG cell wall anchor domain		22-42	98.89
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		102-108	97.70
TIGR03501 GlyGly-CTERM domain		30-42	97.26
TIGR00756 pentatricopeptide repeat domain		102-125	92.98
D2092.1a	*Caenorhabditis elegans*	1-92	92.91
D2092.1b	*Caenorhabditis elegans*	1-123	87.46
Y77E11A.12a	*Caenorhabditis elegans*	2-64	86.20
SaPosin-like Protein family member (spp-14)	*Caenorhabditis elegans*	19-57	81.56
Multiple C2-domains with two transmembrane regions 1 isoform L	*Homo sapiens*	1-132	80.62
Transmembrane protein	*Mycobacterium tuberculosis*	1-61	79.32

SaPosin-like Protein family member (spp-14)	*Caenorhabditis elegans*	19-57	79.01
F55C12.5c	*Caenorhabditis elegans*	11-65	78.50
Mitochondrial ribosomal protein S23		48-55	78.12
Auxilin; four helix bundle, protein binding; NMR	*Bos taurus*	60-112	77.88
Multiple C2-domains with two transmembrane regions 1 isoform S	*Homo sapien*	1-64	76.81
Y77E11A.12b	*Caenorhabditis elegans*	2-64	76.38
EGF-like-domain, multiple 5	*Mus musculus*	23-47	76.05
Vacuolar H ATPase family member (vha-7)	*Caenorhabditis elegans*	21-55	74.03
Oxidoreductase	*Arabidopsis thaliana*	1-47	73.17
PadR-like family transcriptional regulator	*Nostoc punctiforme*	4-55	72.42
C01F6.2	*Caenorhabditis elegans*	5-98	71.70
CG4832-PE, isoform E	*Drosophila melanogaster*	61-122	71.57
"Winged helix" DNA-binding domain (46785) SCOP seed sequence: d1bm9a_		20-64	71.19
GtrA		8-41	69.48
CG33146-PA	*Drosophila melanogaster*	1-64	68.60
ZK353.4	*Caenorhabditis elegans*	13-46	67.88
Proline-rich cyclin A1-interacting protein		132-184	67.61
YbaB-like (82607) SCOP seed sequence: d1j8ba_		107-127	67.13
paREP7	*Pyrobaculum aerophilum*	108-130	65.89
Integral membrane protein	*Streptomyces coelicolor*	28-56	65.74
CG33171-PE, isoform E	*Drosophila melanogaster*	19-104	65.66
Chemokine-like factor superfamily 3 isoform a **(3)**	*Homo sapiens*	1-82	65.45
SPFH_like core domain of the SPFH superfamily		98-134	65.28
V-type ATPase 116 kDa subunit	*Nitrosopumilus maritimus*	10-55	64.89
Isopentenyl pyrophosphate isomerase	*Thermoplasma acidophilum*	47-148	63.71
Transcriptional regulator, PadR-like family	*Eggerthella lenta*	32-55	63.05
Transcriptional regulator, PadR-like family	*Eggerthella lenta*	32-55	63.05
UCP004555		107-127	62.33
Peptidase A24A prepilin type IV	*Candidatus Korarchaeum cryptofilum OPF8*	14-67	62.23
Metal ion binding	*Arabidopsis thaliana*	14-52	61.57
YbaB-like (82607) SCOP seed sequence: d1puga_		108-127	61.33
Vacuolar H ATPase family member (vha-6)	*Caenorhabditis elegans*	1-55	60.37
ZK945.3	*Caenorhabditis elegans*	104-163	59.46

A-type ATP synthase subunit I	*Haloferax volcanii*	9-55	59.20
NADH dehydrogenase I, A subunit	*Neisseria meningitidis*	20-92	59.19
Sensor histidine kinase	*Bartonella henselae*	11-49	58.76
Integral transmembrane protein 2		24-49	58.48
Membrane-associated phospholipid phosphatase	*Methanopyrus kandleri*	1-87	58.13
CG31247-PB, isoform B	*Drosophila melanogaster*	19-56	57.68
CG31247-PC, isoform C	*Drosophila melanogaster*	19-56	57.45
CG31247-PA, isoform A	*Drosophila melanogaster*	19-56	57.22
CG31247-PD, isoform D	*Drosophila melanogaster*	19-56	57.22
Zinc finger, DHHC domain containing 5	*Mus musculus*	14-49	57.19
Multiple C2-domains with two transmembrane regions 2	*Homo sapiens*	1-64	56.00
UNCoordinated family member (unc-32)	*Caenorhabditis elegans*	7-55	55.98
PadR-like family transcriptional regulator	*Thermofilum pendens*	35-70	55.76
BLASTP			
Mucolipin-2	*Echinococcus granulosus*	97-188	0.22
Motif Scan			
Lysine-rich region profile		121-193	9.643
I-TASSER			
HAT1-interacting factor 1	*Saccharomyces cerevisiae*		1.11
HAT1-interacting factor 1	*Saccharomyces cerevisiae*		0.670
Superkiller protein 3	*Saccharomyces cerevisiae*		0.565
SusD-like carbohydrate binding protein	*Bacteroides vulgatus*		0.564
G-protein-signaling modulator 2	*Mus musculus*		0.562
Partner of inscuteable	*Drosophila melanogaster*		0.553
14-3-3 protein	*Cryptosporidium parvum*		0.551
SusD superfamily protein	*Bacteroides vulgatus*		0.548
Predict Protein			
Protein binding		1-2, 45, 67-68	0.548
Nucleus			
Atome2			
Telomerase reverse transcriptase (TEN domain)	*Tetrahymena thermophila*		68.12
Guanine nucleotide exchange factor P115RHOGEF	*Homo sapiens*		59.68
Vinculin	*Gallus gallus*		44.56
Nucleoprotein	*Andes virus*		41.16
rRNA methyltransferase	*Streptomyces*		39.21

	viridochromogenes		
Ubiquinol cytochrome c oxidoreductase	*Gallus gallus*		36.83
Delta-sleep-inducing peptide immunoreactive peptide	*Sus scrofa*		36.46
HIG1 domain family member 1B	*Homo sapiens*		29.71
VicH protein	*Vibrio cholerae*		29.49
Antifreeze protein type 1 analogue	*Pseudopleuronectes americanus*		28.08
Replication terminator protein	*Bacillus subtilis*		25.98
LCoR protein	*Homo sapiens*		22.26
Regulatory protein MNT	*Enterobacteria phage P22*		21.14
Talin-1	*Mus musculus*		18.94
Acetyl-delta-toxin	*Staphylococcus aureus*		16.20

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XX. *Venustaconcha ellipsiformis* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		11-13	99.30
TIGR03304 outer membrane insertion C-terminal signal		25-29	99.16
TIGR01167 LPXTG cell wall anchor domain		19-34	98.97
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		71-78	97.97
TIGR03501 GlyGly-CTERM domain		28-35	97.29
TIGR00756 pentatricopeptide repeat domain		7-25	95.33
ComGC		21-38	92.53
VAMP-5_synaptobrevin		13-44	81.17
Protein-export membrane protein	*Bartonella henselae*	11-87	81.16
Transducer protein Htr37	*Haloferax volcanii*	1-40	79.70
Transducer protein Htr36	*Haloferax volcanii*	3-40	73.90

Stage III sporulation protein AF		7-39	70.15
General secretion pathway protein H	*Nostoc punctiforme*	5-39	69.69
Competence protein ComGC		21-38	69.08
CG11815-PA	*Drosophila melanogaster*	53-75	68.40
C-type LECtin family member (clec-35)	*Caenorhabditis elegans*	13-86	66.92
Protein involved in cis-Golgi membrane traffic	*Saccharomyces cerevisiae*	2-40	66.39
F08F8.8	*Caenorhabditis elegans*	6-40	66.33
Methyl-accepting chemotaxis protein	*Beggiatoa sp. PS*	10-35	65.39
SecD-TM1 SecD export protein N-terminal TM region		12-39	65.33
Vesicle transport through interaction with t-SNAREs 1B homolog	*Mus musculus*	6-40	65.33
d.24.1 Pili subunits (54523) SCOP seed sequence: d2pila_.		20-40	63.63
Vesicle transport through interaction with t-SNAREs 1B	*Homo sapiens*	6-40	63.51
Vesicle-associated membrane protein 5 (myobrevin)	*Homo sapiens*	13-45	61.92
Vesicle transport v-snare protein	*Schizosaccharomyces pombe*	6-40	60.94
Y57G11C.4	*Caenorhabditis elegans*	2-39	60.92
Stage III sporulation protein AF		1-39	60.78
d.24.1 Pili subunits (54523) SCOP seed sequence: d1oqwa_		20-41	59.91
Syntaxin-like t-SNARE	*Saccharomyces cerevisiae*	6-74	59.63
VTI11; receptor	*Arabidopsis thaliana*	6-40	59.48
Related to VTI1 - v-SNARE		6-40	59.10
Methyl-accepting chemotaxis protein II	*Yersinia pestis*	4-40	59.10
CG3279-PA	*Drosophila melanogaster*	6-52	58.98
OapA_N: Opacity-associated protein A N-terminal motif		16-37	58.44
VTI12; SNARE binding / receptor	*Arabidopsis thaliana*	2-40	58.30
SYP123; t-SNARE	*Arabidopsis thaliana*	5-42	57.83
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-34)	*Caenorhabditis elegans*	35-80	54.60
OapA_N Opacity-associated protein A N-terminal motif		16-37	54.42
B0272.2	*Caenorhabditis elegans*	6-40	52.87
DevC protein	*Nostoc punctiforme*	1-42	52.59
Stage III sporulation protein AF		7-39	52.31
F41F3.3	*Caenorhabditis elegans*	18-37	52.22

C-type LECtin family member (clec-25)	*Caenorhabditis elegans*	19-83	51.25
SYP124; t-SNARE	*Arabidopsis thaliana*	14-40	50.99
Integral membrane sensor signal transduction histidine kinase	*Nostoc punctiforme*	5-41	50.79
COLlagen family member (col-144)	*Caenorhabditis elegans*	7-42	50.47
Methyl-accepting chemotaxis protein III	*Escherichia coli*	10-35	50.16
v-SNARE protein involved in Golgi transpor	*Saccharomyces cerevisiae*	2-35	49.57
v-SNARE	*Saccharomyces cerevisiae*	6-40	48.46
Flagellar M-ring protein	*Bacillus subtilis*	2-55	48.18
DevC protein	*Nostoc punctiforme*	3-41	47.10
Resistance to inhibitors of cholinesterase 3 homolog	*Homo sapiens*	17-81	45.89
VTI13; SNARE binding / receptor	*Arabidopsis thaliana*	6-40	45.23
Enzyme inhibitor/ pectinesterase	*Arabidopsis thaliana*	4-83	45.14
Sulfate ABC transporter	*Nostoc punctiforme*	9-35	44.93
TonB family protein	*Nostoc punctiforme*	14-57	43.89
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	14-41	42.86
T24B1.1	*Caenorhabditis elegans*	8-45	42.65
CG4780-PA	*Drosophila melanogaster*	6-42	42.35
COLlagen family member (col-14)	*Caenorhabditis elegans*	1-41	42.28
Related to SNARE protein of Golgi compartment		6-39	42.05
T10E10.5	*Caenorhabditis elegans*	5-42	42.03
DevC protein	*Nostoc punctiforme*	5-41	41.86
FER-1-LIKE		16-43	41.78
Type IV Pilin Pak	*Pseudomonas aeruginosa*	20-39	41.43
Transcriptional accessory factor Tex **(2)**	*Pseudomonas aeruginosa*	55-75	41.42
Syntaxin-related protein required for vacuolar assembly	*Saccharomyces cerevisiae*	5-38	41.06
Methyl-accepting chemotaxis protein	*Bacillus subtilis*	5-41	40.98
CG13581-PA	*Drosophila melanogaster*	77-89	40.49
Diffuse panbronchiolitis critical region 1 protein	*Homo sapiens*	14-56	40.32
Thiol-disulfide oxidoreductase	*Bacillus subtilis*	18-55	39.81
Alpha-disintegrin and metalloproteinase domain 7	*Homo sapiens*	21-84	39.42
Fimbrial protein	*Dichelobacter nodosus*	20-38	39.16
I-TASSER			
Glycine betaine transporter BETP	*Corynebacterium glutamicum*		0.575
Cytochrome P450 130	*Mycobacterium*		0.567

	tuberculosis		
Transcription regulator, Crp family	*Thermus thermophilus*		0.565
DNA topoisomerase 2	*Saccharomyces cerevisiae*		0.556
Virulence-associated V antigen	*Yersinia pestis*		0.554
Vitamin B12 import system permease protein btuC	*Escherichia coli*		0.553
Phase 1 flagellin	*Salmonella enterica*		0.548
Predict Protein			
Protein binding		1, 8, 11, 15, 38, 40-42, 44, 48-50, 66, 70-71, 73	
Mitochondrial membrane			
Atome2			
Hepatitis B virus X-interacting protein	*Homo sapiens*		72.60
Prod 1	*Notophthalmus viridescens*		68.64
Tumor necrosis factor receptor	*Homo sapiens*		63.21
V1V2 region of HIV-1 on 1FD6 scaffold	*Human immunodeficiency virus 1*		57.65
Troponin I, cardiac muscle	*Mus musculus*		38.80
Bone marrow stromal antigen 2	*Homo sapiens*		35.47
Fimbrial protein	*Neisseria gonorrhoeae*		29.39
Photosystem Q(B) protein	*Thermosynechococcus elongatus*		27.56
Fimbrial protein	*Pseudomonas aeruginosa*		23.99
Fimbrial protein	*Dichelobacter nodosus*		23.00
Neurotoxin	*Clostridium botulinum*		13.59

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXI. *Quadrula quadrula* F-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		1-4	99.37
TIGR03304 outer membrane insertion C-terminal signal		16-20	99.21
TIGR01167 LPXTG cell wall anchor domain		10-25	99.02
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		62-69	97.98
TIGR03501 GlyGly-CTERM domain		19-26	97.38
TIGR00756 pentatricopeptide repeat domain		63-66	94.93
ComGC		12-29	94.12
VAMP-5_synaptobrevin		4-35	87.41
Methyl-accepting chemotaxis protein	*Beggiatoa sp. PS*	1-26	84.49
Protein-export membrane protein	*Bartonella henselae*	2-77	83.21
C-type LECtin family member (clec-35)	*Caenorhabditis elegans*	4-77	79.44
ComGC Competence protein ComGC		12-29	77.26
SecD-TM1 SecD export protein N-terminal TM region		3-30	77.09
Pili subunits (54523) SCOP seed sequence: d2pila_		11-30	73.77
Vesicle-associated membrane protein 5 (myobrevin)	*Homo sapiens*	4-36	73.53
Methyl-accepting chemotaxis sensory transducer	*Beggiatoa sp. PS*	1-26	71.68
d.24.1 Pili subunits (54523) SCOP seed sequence: d1oqwa_		11-32	70.76
Methyl-accepting chemotaxis protein III	*Escherichia coli*	1-26	70.47
CG11815-PA	*Drosophila melanogaster*	44-66	69.18
Syntaxin-like t-SNARE	*Saccharomyces cerevisiae*	7-65	68.43
C-type LECtin family member (clec-25)	*Caenorhabditis elegans*	10-74	67.54
SYP123; t-SNARE	*Arabidopsis thaliana*	5-33	67.07
SYP124; t-SNARE	*Arabidopsis thaliana*	5-31	66.45
OapA_N: Opacity-associated protein A N-terminal motif		7-28	65.18
Sensor signal transduction histidine kinase	*Beggiatoa sp. PS*	1-31	63.30
Diffuse panbronchiolitis critical region 1 protein	*Homo sapiens*	5-47	62.48
Spore_III_AF: Stage III sporulation protein AF (Spore_III_AF)		2-30	62.09
OapA_N Opacity-associated protein A N-terminal motif		7-28	61.64
F41F3.3	*Caenorhabditis elegans*	9-28	59.91
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	5-32	57.86

Plasma membrane t-SNARE	*Saccharomyces cerevisiae*	7-26	56.77
FER-1-LIKE		7-34	56.30
Methyl-accepting chemotaxis sensory transducer	*Beggiatoa sp. PS*	1-26	55.82
Cytochrome c family protein	*Beggiatoa sp. PS*	1-34	55.53
Vesicle transport through interaction with t-SNAREs 1B	*Homo sapiens*	3-31	55.19
CreD		3-24	55.15
Vesicle transport through interaction with t-SNAREs 1B homolog	*Mus musculus*	3-31	54.58
RCR		9-26	54.56
Pilin PilE	*Neisseria meningitidis*	1-29	54.48
Vesicle-associated membrane protein 1 isoform 1	*Homo sapiens*	4-31	54.42
Methyl-accepting chemotaxis protein II	*Yersinia pestis*	1-32	54.19
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-34)	*Caenorhabditis elegans*	26-71	54.00
C-type LECtin family member (clec-32)	*Caenorhabditis elegans*	3-74	53.81
Type IV Pilin Pak	*Pseudomonas aeruginosa*	11-30	53.69
F08F8.8	*Caenorhabditis elegans*	3-31	53.66
Stage III sporulation protein AF		2-30	53.25
Stage III sporulation protein AG		7-25	53.24
Sensor protein	*Nostoc punctiforme*	3-32	52.91
Protein involved in cis-Golgi membrane traffic; v-SNARE	*Saccharomyces cerevisiae*	3-31	52.91
COLlagen family member (col-176)	*Caenorhabditis elegans*	3-33	52.78
Syntaxin-related protein required for vacuolar assembly	*Saccharomyces cerevisiae*	5-31	52.32
C-type LECtin family member (clec-38)	*Caenorhabditis elegans*	10-80	52.29
TonB family protein	*Nostoc punctiforme*	5-48	51.68
Extracellular solute-binding protein	*Thermofilum pendens*	3-26	51.66
USE1-like protein		9-26	51.50
Thiol-disulfide oxidoreductase	*Bacillus subtilis*	9-46	51.47
Fimbrial protein (cell adhesion)	*Dichelobacter nodosus*	11-29	51.31
C-type LECtin family member (clec-27)	*Caenorhabditis elegans*	10-74	51.13
Enzyme inhibitor/ pectinesterase	*Arabidopsis thaliana*	4-74	51.02
Resistance to inhibitors of cholinesterase 3 homolog	*Homo sapiens*	8-72	50.86
Sialidase	*Haloferax volcanii*	3-26	50.72
Syntaxin 12	*Mus musculus*	5-30	50.68
Use1 Membrane fusion protein Use1		9-26	50.43
Target membrane receptor (t-SNARE)	*Saccharomyces cerevisiae*	7-26	50.43

CG3279-PA	*Drosophila melanogaster*	3-31	50.13
Fimbrial protein	*Dichelobacter nodosus*	11-29	49.97
Plasma membrane t-SNARE	*Saccharomyces cerevisiae*	6-26	49.94
SYP111; t-SNARE	*Arabidopsis thaliana*	5-34	49.48
Sensor protein	*Nostoc punctiforme*	3-32	49.45
Y57G11C.4	*Caenorhabditis elegans*	3-30	48.87
VTI11; receptor	*Arabidopsis thaliana*	3-31	48.51
Alpha-disintegrin and metalloproteinase domain 7	*Homo sapiens*	12-75	48.44
Use1: Membrane fusion protein Use1		9-26	48.26
SYP21; t-SNARE	*Arabidopsis thaliana*	5-30	47.91
4HB_MCP_1: Four helix bundle sensory module for signal transduction		7-26	47.71
CG31136-PA	*Drosophila melanogaster*	7-32	47.53
PSIBLAST			
Preprotein translocase subunit SecG	*Bergeyella zoohelcum*	2-80	2.00e-04
Leucine rich repeat protein **(2)**	*Leptospira kirschneri*	19-63	3.00e-04, 7.00e-04
Magnesium transporter MgtE **(2)**	*Thermus oshimai*	7-49	7.00e-04
Mg2+ transporter MgtE	*Thermus oshimai*	7-49	7.00e-04
Histidine kinase	*Paenibacillus larvae*	1-26	8.00e-04
Peptidase M15B and M15C DD-carboxypeptidase VanY/endolysin	*Paenibacillus sp. JDR-2*	13-74	9.00e-04
Histidine kinase	*Paenibacillus larvae*	1-26	0.001
Peptidase M15	*Paenibacillus sp. JDR-2*	13-74	0.001
Calcium/proton exchanger **(3)**	*Cryptococcus gattii*	26-78	0.001
Calcium ion transporter	*Cryptococcus gattii*	26-78	0.001
Diaminopimelate epimerase	*Pseudomonas sp. RIT357*	29-78	0.001
Amino acid transporter	*Olleya marilimosa*	14-49	0.002
Sulfate transporter	*Bacillus cereus*	3-37	0.002
Serine/threonine protein kinase	*Vibrio harveyi*	6-67	0.002
Na(+)/H(+) antiporter NhaA	*Salinispora pacifica*	13-75	0.002
C4-dicarboxylate ABC transporter	*Chelatococcus sp. GW1*	7-38	0.002
Poly(glycerophosphate chain) D-alanine transfer protein	*Streptococcus parauberis*	7-61	0.002
D-alanyl-lipoteichoic acid biosynthesis protein DltD	*Streptococcus parauberis*	7-61	0.002
Sodium:proton antiporter	*Salinispora pacifica*	13-75	0.002
ATP-dependent DNA helicase RecQ	*Rhodopirellula europaea*	37-76	0.002

Transporter, MotA/TolQ/ExbB proton channel family protein	*Prevotella pleuritidis*	11-78	0.003
Sulfate transporter	*Bacillus cereus*	3-40	0.003
Bacterial membrane protein YfhO	*Microvirga lotononidis*	5-73	0.003
MULTISPECIES: C4-dicarboxylate ABC transporter	*Rhizobium*	7-38	0.003
XRE family transcriptional regulator	*Cyanothece sp. PCC 8801*	23-49	0.003
Leucine-rich repeat and death domain-containing protein	*Heterocephalus glaber*	19-59	0.003
Serine/threonine protein kinase **(2)**	*Vibrio harveyi*	6-67	0.003, 0.004
Membrane protein, partial	*Streptomyces xanthophaeus*	6-34	0.004
Doublesex-and mab-3-related transcription factor 3	*Strongyloides ratti*	35-66	0.004
Membrane protein, partial	*Streptomyces xanthophaeus*	6-34	0.004
General secretion pathway protein G	*Gallaecimonas xiamenensis*	12-57	0.004
Nicotinate (nicotinamide) nucleotide adenylyltransferase	*Cryptococcus gattii*	27-74	0.004
Bicarbonate transporter BicA	*Prochlorococcus marinus*	10-61	0.004
CoA-binding protein	*Halosarcina pallida*	34-74	0.004
Antisigma-factor antagonist, STAS	*Bacillus cereus Rock3-28*	3-40	0.005
Magnesium transporter MgtE	*Thermus yunnanensis*	10-49	0.005
I-TASSER			
A-type ATP synthase subunit E	*Methanocaldococcus jannaschii*		1.18
Fumarase C	*Escherichia coli*		0.548
Fumarate hydratase	*Homo sapiens*		0.543
Deoxyguanosinetriphosphate triphosphohydrolase	*Escherichia coli K-1*		0.543
Fumarate hydratase class II	*Rickettsia prowazekii*		0.542
Adenylosuccinate lyase	*Staphylococcus aureus*		0.538
Fumarate lyase	*Chelativorans sp. BNC1*		0.538
Fumarase Fum	*Mycobacterium marinum M*		0.537
Adenylosuccinate lyase	*Bacillus anthracis*		0.537
Predict Protein			
Protien binding		1, 13, 39-40, 59, 65,	

		85	
Mitochondrial membrane			
Atome2			
HIV-1 matrix protein **(2)**	*Human immunodeficiency virus 1*		83.13, 72.79
Obscurin-like protein 1	*Homo sapiens*		41.74
Photosystem Q(B) protein **(2)**	*Thermosynechococcus elongatus*		37.59, 31.35
Photosystem II reaction center protein T	*Mastigocladus laminosus*		27.60
Sec-independent protein translocase protein tatAd	*Bacillus subtilis*		27.54
Fimbrial protein	*Neisseria gonorrhoeae*		26.93
Fimbrial protein	*Pseudomonas aeruginosa*		23.40
Vesicle-associated membrane protein 2	*Rattus norvegicus*		23.05
Fimbrial protein	*Dichelobacter nodosus*		22.81
Defensin	*Caretta caretta*		16.71

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXII. *Pyganodon grandis* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR01167 LPXTG cell wall anchor domain		62-67	99.45
TIGR03304 outer membrane insertion C-terminal signal		1-6	99.34
TIGR04294 prepilin-type processing-associated H-X9-DG domain		26-29	99.27
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		54-61	97.97
TIGR03501 GlyGly-CTERM domain		16-26	97.22
TIGR00756 pentatricopeptide repeat domain		46-53	94.27
CG7685-PA	*Drosophila melanogaster*	7-34	92.74
Intra-Golgi v-SNARE	*Saccharomyces cerevisiae*	2-34	89.49
TMEM156: TMEM156 protein family		9-35	76.75,

			76.59
Conserved inner membrane protein	*Escherichia coli*	5-37	75.88
SrtB		4-42	75.38
LptF_YjgP LPS export ABC transporter permease LptF		7-37	74.47
GOLGI SNARE BET1-RELATED		11-32	69.38
Ceramidase		11-52	69.27
CG13969-PA	*Drosophila melanogaster*	11-52	68.78
Sensor histidine kinase	*Streptococcus pneumoniae*	9-48	64.48
LPS export ABC transporter permease LptG		5-37	64.01
Sensory box histidine kinase PhoR	*Staphylococcus aureus*	9-44	63.97
Nitric oxide reductase subunit C	*Pseudomonas aeruginosa*	5-30	61.75
Saliv_gland_allergen_Aed3		10-27	60.13
Essential SNARE protein localized to the ER	*Saccharomyces cerevisiae*	13-34	60.06
T27F7.3a	*Caenorhabditis elegans*	10-44	59.65
CG11020-PA, isoform A	*Drosophila melanogaster*	5-46	57.90
Sterol reductase/lamin b receptor		27-55	57.55
Alkaline ceramidase 2	*Mus musculus*	11-52	56.62
Dipeptide transport permease	*Pyrobaculum aerophilum*	2-34	56.60
GRP: Glycine rich protein family		14-34	56.22
Permease YjgP/YjgQ family protein	*Nostoc punctiforme*	5-37	56.02
Osm-9 & capsaicin receptor-related family (ocr-4)	*Caenorhabditis elegans*	5-70	55.92
ATBS14A; protein transporter	*Arabidopsis thaliana*	11-32	55.52
W02F12.2	*Caenorhabditis elegans*	11-61	54.80
Related to YPC1 - Alkaline ceramidase		11-52	53.56
Human EMeRin homolog family member (emr-1)	*Caenorhabditis elegans*	13-31	53.29
GDSL family lipase	*Nitrosopumilus maritimus*	1-38	52.36
Alkaline ceramidase that also has reverse (CoA-independent) ceramide synthase function	*Saccharomyces cerevisiae*	7-52	52.01
Vesicle-associate membrane protein-associated protein		12-34	51.58
SVM protein signal sequence		10-31	51.26
GRP Glycine rich protein family		13-34	51.08
CbiN ABC-type cobalt transport system, periplasmic component		9-44	49.52
Peptidoglycan-associated lipoprotein Pal	*Yersinia pestis*	8-27	47.95
Lipoprotein required for capsular polysaccharide translocation through the outer membrane	*Escherichia coli*	9-27	47.20

Retinoblastoma-associated protein	*Homo sapiens*	37-54	47.14
Galactose-3-O-sulfotransferase 3	*Homo sapiens*	4-63	46.93
H/K_exch_ATPase_C		9-38	46.76
Phytoceramidase, alkaline	*Homo sapiens*	2-52	46.21
DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	9-49	46.06
Y41C4A.19	*Caenorhabditis elegans*	2-54	45.83
Secreted protein	*Beggiatoa sp. PS*	9-57	45.68
Protein transporter	*Arabidopsis thaliana*	11-34	44.80
Extracellular solute-binding protein	*Nostoc punctiforme*	2-31	43.64
I-TASSER			
Human cyclin B1	*Homo sapiens*		0.520
G1/S-specific cyclin-D1	*Homo sapiens*		0.518, 0.512
V-cyclin	*Human herpesvirus 8*		0.517
G1/S-specific cyclin E1	*Homo sapiens*		0.517
Cell division protein kinase 4	*Homo sapiens*		0.516
Cell division protein kinase 2	*Homo sapiens*		0.515
G2/mitotic-specific cyclin-B1	*Homo sapiens*		0.513
Cyclin-C	*Homo sapiens*		0.513
Predict Protein			
Protein binding		1-6, 9, 28-29, 31-33, 35, 38-40, 42, 57, 59, 61-62, 64	
Mitochondrial membrane			
Atome2			
Colicin D (colicin D catalytic domain)	*Escherichia coli*		87.77
14 kDa phosphohistidine phosphatase	*Homo sapiens*		83.70
Thrombin	*Homo sapiens*		67.65
Human beta2-Glycoprotein I	*Homo sapiens*		66.28
Calcium-gated potassium channel mthK	*Methanothermobacter thermautotrophicus*		48.14

Photosynthetic reaction center C subunit	*Thermochromatium tepidum*		38.65
CREB-binding protein	*Mus musculus*		37.60
Transcription factor Dp-1	*Homo sapiens*		36.50
CPAP	*Danio rerio*		27.81
Vesicle-associated membrane protein 2	*Rattus norvegicus*		24.76
Protein translocase subunit secA	*Bacillus subtilis*		24.71
Transient receptor potential cation channel subfamily V member 1	*Rattus norvegicus*		20.75
Antibody fab fragment light chain	*Mus musculus*		

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXIII. *Inversidens japanensis* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR01167 LPXTG cell wall anchor domain		1-15	99.44
TIGR03304 outer membrane insertion C-terminal signal		3-5	99.36
TIGR04294 prepilin-type processing-associated H-X9-DG domain		13-15	99.23
X-X-X-Leu-X-X-Gly heptad repeats		57-65	97.91
TIGR03501 GlyGly-CTERM domain		2-13	97.58
TIGR00756 pentatricopeptide repeat domain		14-18	94.24
L-lactate permease-related protein	*Neisseria meningitidis*	1-31	81.61
golgi phosphoprotein 4	*Homo sapiens*	4-18	80.40
H/K_exch_ATPase_C		3-15	80.09
Md_memb_hyd		5-24	74.97
Cell division protein ZipA	*Yersinia pestis*	2-18	72.32
Cell division protein ZipA	*Escherichia coli*	2-17	71.55
Cell division protein ZipA	*Pseudomonas aeruginosa*	2-18	70.81
C-type LECtin family member (clec-62)	*Caenorhabditis elegans*	1-19	70.72,

			69.95
TGF-beta-activated kinase 1 and MAP3K7-binding PR	*Homo sapiens*	40-59	69.82
ABI gene family, member 3 (NESH) binding protein	*Homo sapiens*	1-14	69.57
Asialoglycoprotein receptor 1	*Homo sapiens*	1-14	69.16
Melanocortin 3 receptor	*Homo sapiens*	16-61	67.54
Fibronectin-binding_protein_I_partial TQXA domain		10-22	66.79
T10G3.1	*Caenorhabditis elegans*	2-13	65.01
Y111B2A.26	*Caenorhabditis elegans*	17-51	64.36
Alpha-1,4-N-acetylglucosaminyltransferase	*Homo sapiens*	1-15	61.44
TIGR03778 VPDSG-CTERM protein sorting domain		56-61	65.10
CG33206-PB, isoform B	*Drosophila melanogaster*	7-60	60.26
Y41G9A.4a	*Caenorhabditis elegans*	1-32	60.13
Thrombospondin type 3 repeat-containing protein	*Nitrosopumilus maritimus*	2-27	60.12
Cell division protein ZipA		1-21	59.36
Neuropeptide-Like Protein family member (nlp-16)	*Caenorhabditis elegans*	1-15	59.23
Peptidyl-prolyl cis-trans isomerase	*Neisseria meningitidis*	3-60	58.90
CG12522-PA	*Drosophila melanogaster*	2-51	58.86
Cell division protein	*Yersinia pestis CO92*	1-51	57.29
TMEM52: Transmembrane 52		2-13	56.98
ZK1010.5	*Caenorhabditis elegans*	2-16	56.32
C-type lectin, superfamily member 14 isoform 2	*Homo sapiens*	2-18	56.21
KdpC K+-transporting ATPase, c chain		1-15	55.34
Submaxillary gland androgen regulated protein 1	*Mus musculus*	1-11	54.45
Transmembrane protein	*Mycobacterium tuberculosis*	1-40	54.33
Potassium-transporting ATPase subunit C; Reviewed		1-15	54.00
Alpha 1B-glycoprotein	*Homo sapiens*	1-33	53.56
KdpC: K+-transporting ATPase, c chain		3-15	53.51
CG33706-PA, isoform A	*Drosophila melanogaster*	2-14	52.88
Potassium-transporting ATPase C chain K+		1-15	52.85
Potassium-transporting ATPase subunit C	*Nostoc punctiforme*	3-15	52.73
F13G3.12	*Caenorhabditis elegans*	13-19	52.66
CG11709-PA	*Drosophila melanogaster*	1-55	52.32
Potassium-transporting ATPase subunit C	*Escherichia coli*	2-15	51.89
Potassium-transporting ATPase subunit C	*Nostoc punctiforme*	2-15	51.32
Potassium-transporting ATPase subunit C	*Mycobacterium*	1-15	51.31

	tuberculosis		
Golgi phosphoprotein 4	*Mus musculus*	4-18	51.08
ROK family transcriptional regulator	*Streptomyces coelicolor*	2-35	50.72
Cell envelope integrity inner membrane protein TolA	*Yersinia pestis*	3-26	50.33
Chymotrypsinogen B2	*Homo sapiens*	1-29	50.13
Macrophage galactose N-acetyl-galactosamine specific lectin 2	*Mus musculus*	1-17	49.30
CG9928-PA	*Drosophila melanogaster*	1-15	49.29
Potassium-transporting ATPase subunit C	*Yersinia pestis*	2-15	48.97
Ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin)	*Homo sapiens*	5-25	48.15
K+-transporting ATPase, C subunit	*Staphylococcus aureus*	2-15	47.95
R09D1.5	*Caenorhabditis elegans*	2-19	47.79
CD8 ALPHA CHAIN		2-13	47.69
P-type ATPase	*Frankia alni*	2-15	47.63
GRP: Glycine rich protein family		1-52	47.53
Outer membrane efflux protein	*Nostoc punctiforme*	1-50	47.40
Potassium-transporting ATPase subunit C	*Pseudomonas aeruginosa*	2-15	47.35
Cation transport system component	*Streptomyces coelicolor*	3-15	46.51
I-TASSER			
MSin3A-binding protein	*Mus musculus*		1.03
Predict Protein			
Protein binding		1-4, 15-19, 28, 30-33, 35, 45-47, 49, 53-54, 56-58, 62, 65	
Secreted			
Atome2			
Carboxypeptidase A2	*Homo sapiens*		69.57
TraF protein	*Escherichia coli*		61.48
Nonstructural RNA-binding protien 34	*Simian rotavirus*		28.60-48.04

Nucleoporin	*Mus musculus*		47.00
Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 4	*Homo sapiens*		41.53
Chromo domain-containing protein 1	*Saccharomyces cerevisiae*		38.89
Bone marrow stromal antigen 2	*Homo sapiens*		37.80
Cytochrome b6 **(3)**	*Mastigocladus laminosus*		29.46-34.43
Presenilin-1	*Homo sapiens*		32.35
Polyadenylate-binding protein 1	*Homo sapiens*		28.20
Mitogen-activated protein kinase 14	*Mus musculus*		25.10

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXIV. *Utterbackia peninsularis* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR01167 LPXTG cell wall anchor domain		47-52	99.47
TIGR04294 prepilin-type processing-associated H-X9-DG domain		18-21	99.34
TIGR03304 outer membrane insertion C-terminal signal		27-28	99.28
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		39-46	98.08
TIGR03501 GlyGly-CTERM domain		8-18	97.39
TIGR00756 pentatricopeptide repeat domain		31-38	94.79
CG7685-PA	*Drosophila melanogaster*	2-34	90.11
CG11786-PA	*Drosophila melanogaster*	1-33	83.58
CbiN ABC-type cobalt transport system, periplasmic component		1-29	79.63
Human EMeRin homolog family member (emr-1)	*Caenorhabditis elegans*	5-24	74.62
CG13969-PA	*Drosophila melanogaster*	3-37	74.59
Ceramidase		3-37	67.66
Saliv_gland_allergen_Aed3		2-19	66.83

GRP: Glycine rich protein family		6-23	64.89
Lipoprotein required for capsular polysaccharide translocation through the outer membrane	*Escherichia coli*	1-19	62.25
Intra-Golgi v-SNARE	*Saccharomyces cerevisiae*	3-23	57.13
Syntaxin-like t-SNARE	*Saccharomyces cerevisiae*	8-58	56.27
Syntaxin 5	*Mus musculus*	3-25	56.06
Cysteine-type endopeptidase/ cysteine-type peptidase	*Arabidopsis thaliana*	2-41	55.67
GRP Glycine rich protein family		5-23	55.06
T27F7.3a	*Caenorhabditis elegans*	2-29	54.72
RCR		8-23	54.67
W02F12.2	*Caenorhabditis elegans*	3-46	54.24
CG4214-PA, isoform A	*Drosophila melanogaster*	6-25	54.12
CG4214-PB, isoform B	*Drosophila melanogaster*	6-25	54.12
SYP31; t-SNARE	*Arabidopsis thaliana*	3-24	53.28
Retinoblastoma-associated protein	*Homo sapiens*	22-39	53.01
Alkaline ceramidase 2	*Mus musculus*	3-37	52.35
MORN repeat protein	*Beggiatoa sp. PS*	1-21	52.26
Cancer susceptibility candidate 4 isoform 1	*Mus musculus*	4-32	52.12
Golgi phosphoprotein 2	*Homo sapiens*	1-32	52.03
Golgi phosphoprotein 2	*Homo sapiens*	1-32	52.03
Target membrane receptor (t-SNARE)	*Saccharomyces cerevisiae*	3-24	51.69
Sensor histidine kinase	*Streptococcus pneumoniae*	1-35	51.24
SYNtaxin family member (syn-3)	*Caenorhabditis elegans*	8-25	51.20
C46H11.8	*Caenorhabditis elegans*	6-20	50.34
Golgi SNARE BET1-related		3-23	49.96
Sensory box histidine kinase PhoR	*Staphylococcus aureus*	1-35	49.94
Related to YPC1 - Alkaline ceramidase		3-37	49.81
P53-induced protein related		4-33	48.53
LCR32	*Arabidopsis thaliana*	1-18	48.29
Sensor protein	*Nostoc punctiforme*	2-35	47.97
Cancer susceptibility candidate 4 isoform b	*Homo sapiens*	4-32	47.39
SVM protein signal sequence		2-21	46.80
Cancer susceptibility candidate 4 isoform a	*Homo sapiens*	4-32	46.01
Rhodanese-like protein	*Beggiatoa sp. PS*	2-21	45.74
LCR9	*Arabidopsis thaliana*	1-18	45.22
Cell wall structural complex MreBCD transmembrane	*Escherichia coli*	2-34	45.03

component MreC			
Cancer susceptibility candidate 4 isoform 2	*Mus musculus*	4-32	44.69
Alkaline ceramidase that also has reverse (CoA-independent) ceramide synthase activity	*Saccharomyces cerevisiae*	3-37	44.61
T01B10.5	*Caenorhabditis elegans*	9-67	44.41
Pectinesterase/pectinesterase inhibitor	*Arabidopsis thaliana*	2-21	44.24
Conserved inner membrane protein	*Escherichia coli*	1-32	44.17
Y116A8C.44	*Caenorhabditis elegans*	10-21	44.15
SYP32; t-SNARE	*Arabidopsis thaliana*	3-25	44.03
v-SNARE protein involved in Golgi transport, homolog of the mammalian protein GOS-28/GS28	*Saccharomyces cerevisiae*	7-29	43.63
T19H12.3	*Caenorhabditis elegans*	6-21	43.56
F08E10.7	*Caenorhabditis elegans*	6-34	43.53
Alpha/beta hydrolase superfamily protein	*Lactobacillus casei*	1-30	43.32
CG4716-PB, isoform B	*Drosophila melanogaster*	10-35	43.04
F10B5.9	*Caenorhabditis elegans*	9-33	42.75
Leukocyte surface antigen CD47		2-34	42.10
COLlagen family member (col-102)	*Caenorhabditis elegans*	4-39	41.52
Diguanylate cyclase	*Nostoc punctiforme*	2-35	41.51
F58G1.5	*Caenorhabditis elegans*	8-55	41.39
Predict Protein			
Protein binding		1-2, 41, 50	
Mitochondrial membrane			
Atome2			
Thrombin	*Homo sapiens*		86.79
Spindle pole body component SPC42	*Saccharomyces cerevisiae*		67.12
Antitoxin RelB3	*Methanocaldococcus jannaschii*		62.24
Antifreeze peptide SS-3	*Myoxocephalus scorpius*		55.05
CREB-binding protein	*Mus musculus*		50.10
Antifreeze peptide SS-3	*Myoxocephalus scorpius*		45.83
Oligomerization	*Homo sapiens*		45.03
BM2 protein	*Influenza B virus*		42.29
Importin subunit alpha-2	*Mus musculus*		41.49
Protein transport protein SEC23	*Saccharomyces cerevisiae*		41.12

CPAP	*Danio rerio*		40.65
Transcription factor Dp-1	*Homo sapiens*		36.18
Protein translocase subunit secA	*Bacillus subtilis*		23.61
Argenine attenuator peptide	*Neurospora crassa*		23.04
Beta-hemolysin	*Staphylococcus aureus*		13.11

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXV. *Solenaia carinatus* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		2-9	99.37
TIGR03304 outer membrane insertion C-terminal signal		62-63	99.06
TIGR01167 LPXTG cell wall anchor domain		4-22	98.91
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		4-7	97.69
TIGR03501 GlyGly-CTERM domain		7-17	96.90
TIGR00756 pentatricopeptide repeat domain		16-23	93.28
5-hydroxytryptamine (serotonin) receptor 1D	*Homo sapiens*	33-70	92.61
Light-harvesting complex subunits	*Rhodoblastus acidophilus*	8-26	87.30, 87.07
Light-harvesting complex subunits (56918) SCOP seed sequence: d1dx7a_		4-26	83.17
Light-harvesting complex subunits (56918) SCOP seed sequence: d1lghb_		4-26	81.89
Light-harvesting complex subunits (56918) SCOP seed sequence: d1kzub_		4-26	80.77
Prenylated RAB acceptor 1-related		6-26	79.07
Light-harvesting protein B-800/850		5-26	74.63
Integrin, beta-like 1	*Mus musculus*	1-28	73.53
Light-harvesting complex subunits	*Rhodoblastus acidophilus*	8-25	72.83

Light-harvesting complex subunits (56918) SCOP seed sequence: d1ijdb_		4-25	71.44
LH1 beta polypeptide; photosynthesis		5-26	70.95
Light-harvesting complex subunits	*Rhodoblastus acidophilus*	8-25	70.88
P-loop containing nucleoside triphosphate hydrolases (52540) SCOP seed sequence: d1qhla_		48-56	70.54
LH II, B800/850, light harvesting complex II		5-26	69.61
Rab acceptor 1	*Homo sapiens*	6-26	69.24
Membrane protein	*Beggiatoa sp. PS*	34-54	68.30
Rab acceptor 1	*Mus musculus*	6-26	67.03
Transmembrane protein HTP-1 related		2-21	66.89
CG1418-PA	*Drosophila melanogaster*	6-26	65.31
Dienelactone hydrolase	*Nostoc punctiforme*	8-66	57.49
PRA1 PRA1 family protein		6-26	57.44
ZK896.1	*Caenorhabditis elegans*	4-47	57.43
Sterol carrier protein 2 isoform 3 precursor	*Homo sapiens*	23-31	56.70
MPI7	*Arabidopsis thaliana*	6-26	56.39
Syntaxin-like t-SNARE	*Saccharomyces cerevisiae*	8-66	56.24
Protein localized to COPII vesicles	*Saccharomyces cerevisiae*	6-26	55.8
LH-1, light-harvesting protein B-880, beta chain	*Rhodospirillum rubrum*	5-26	55.18
Magnesium transporter	*Synechococcus sp. CC9311*	9-28	55.18
Flagellar motor protein MotS		11-23	55.17
CG10031-PA	*Drosophila melanogaster*	1-21	54.86
Integrin, beta-like 1 (with EGF-like repeat domains)	*Homo sapiens*	1-24	54.45
Phosphatidylserine decarboxylase	*Methanopyrus kandleri*	1-26	53.64
Light-harvesting protein B-880, beta chain		5-26	53.58
CG6339-PA	*Drosophila melanogaster*	48-56	52.81
Cell surface glycoprotein	*Methanosarcina mazei*	1-63	51.14
Excinuclease ATPase subunit	*Beggiatoa sp. PS*	48-56	50.17
Dopamine receptor D1A	*Mus musculus*	1-26	50.08
RADiation sensitivity abnormal/yeast RAD-related family member (rad-50)	*Caenorhabditis elegans*	34-56	49.33
Light-harvesting protein B-880, beta chain	*Rhodospirillum rubrum*	5-26	49.09
P-loop containing nucleoside triphosphate hydrolases (52540) SCOP seed sequence: d1np6a_		48-56	48.83

RAD50; ATP binding / nuclease/ zinc ion binding	*Arabidopsis thaliana*	48-56	48.65
ABC transporter ATP-binding protein	*Beggiatoa sp. PS*	48-56	48.37
P-loop containing nucleoside triphosphate hydrolases (52540) SCOP seed sequence: d1q3ta_		48-56	48.27
P-loop containing nucleoside triphosphate hydrolases (52540) SCOP seed sequence: d1f2t.1		48-56	48.11
Oligosaccharyltransferase subunit ost4p	*Saccharomyces cerevisiae*	7-21	47.89
Oligosaccharyltransferase subunit ost4p	*Saccharomyces cerevisiae*	7-21	47.89
Subunit of MRX complex	*Saccharomyces cerevisiae*	48-56	47.83
Glutamine ABC transporter (glutamine-binding protein)	*Bacillus subtilis*	5-64	47.75
UCP018933		47-69	47.22
P-loop containing nucleoside triphosphate hydrolases (52540) SCOP seed sequence: d1gkya_		48-56	47.06
ABC transporter	*Beggiatoa sp. PS*	48-56	46.29
LHC Antenna complex alpha/beta subunit		5-26	46.13
P-loop containing nucleoside triphosphate hydrolases (52540) SCOP seed sequence: d1e3ma2		48-56	45.70
RAD50 homolog isoform 1	*Homo sapiens*	48-56	45.17
ATP-binding protein	*Beggiatoa sp. PS*	48-56	44.98
MotB flagellar motor protein MotB		11-23	44.88
F57C2.5	*Caenorhabditis elegans*	4-38	44.66
Glycine rich protein family		1-53	44.60
UbiA prenyltransferase	*Nostoc punctiforme*	9-31	44.58
yidC_nterm membrane protein insertase, YidC/Oxa1 family, N-terminal domain		7-77	44.52
BLASTP			
Bifunctional 2',3'-cyclic nucleotide 2'-phosphodiesterase/3'-nucleotidase precursor protein		12-81	2.98e-03
I-TASSER			
Pilin, type IV	*Thermus thermophilus*		1.04, 1.07
Anastral spindle 2, SAS 4	*Drosophila melanogaster*		1.00
Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	*Homo sapiens*		0.617
Fructose 1,6-bisphosphatase/inositol monophosphatase	*Archaeoglobus fulgidus*		0.617
Inositol monophosphatase	*Zymomonas mobilis*		0.606
Pilin, type IV	*Thermus thermophilus*		0.591

Pseudopilin GspK	*Escherichia coli*		0.589
Fructose 1,6-bisphosphatase	*Pisum sativum*		0.589
Fimbrial protein	*Pseudomonas aeruginosa*		0.588
Xaa-Pro aminopeptidase 1	*Homo sapiens*		0.588
Type IV pilin	*Pseudomonas aeruginosa*		0.580
Predict Protein			
Protein binding		1-2, 22, 25, 28, 32, 34-37, 57, 80, 85-86	
Secreted			
Atome2			
Major capsid protein (protein P3)	*Enterobacteria phage*		80.01
Importin alpha-1 subunit	*Homo sapiens*		71.81
Type II restriction enzyme HindIII	*Haemophilus influenzae*		66.46
AS-48 protein	*Enterococcus faecalis*		63.35
Stromal cell-derived factor 1	*Homo sapiens*		55.61
Photosynthetic reaction center C subunit	*Thermochromatium tepidum*		49.31
Archaeal adhesion filament core	*Ignicoccus hospitalis*		45.48
Light-harvesting protein B-800/850, alpha chain	*Rhodoblastus acidophilus*		42.22
Light-harvesting protein B-880, beta chain	*Rhodospirillum rubrum*		37.37
Chromosome segregation protein smc	*Pyrococcus furiosus*		31.93
Phosphate starvation-inducible protein	*Corynebacterium glutamicum*		31.63
Guanylate kinase	*Coxiella burnetii*		31.61
Light harvesting complex II	*Phaeospirillum molischianum*		31.23
Chromosome segregation SMC protein	*Thermotoga maritima*		30.86
Cytochrome c oxidase, cbb3-type, subunit N	*Pseudomonas stutzeri*		27.90
Guanylate kinase	*Mus musculus*		27.75
Fructokinase	*Ruegeria sp. TM1040*		27.69

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict

Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXVI. *Cumberlandia monodonta* F-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		12-14	99.21
TIGR04294 prepilin-type processing-associated H-X9-DG domain		48-50	99.04
TIGR01167 LPXTG cell wall anchor domain		72-76	98.83
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		18-27	97.70
TIGR03501 GlyGly-CTERM domain		4-15	96.86
TIGR00756 pentatricopeptide repeat domain		59-68	93.47
F46H6.5	*Caenorhabditis elegans*	56-75	76.05
UCP029505		6-16	69.88
PEP-CTERM protein-sorting domain		64-69	46.91
TIGR04288 CGP-CTERM domain		2-12	46.76
Conserved inner membrane protein	*Escherichia coli*	2-17	40.08
COX7		2-16	36.69
Conserved integral membrane protein	*Corynebacterium diphtheriae*	3-28	34.75
Y54E10BL.2	*Caenorhabditis elegans*	1-27	34.23
Homodimeric domain of signal transducing histidine kinase (47384) SCOP seed sequence: d1joya_		36-42	33.23
DumPY : shorter than wild-type family member (dpy-14)	*Caenorhabditis elegans*	1-27	31.73
T10E10.2	*Caenorhabditis elegans*	1-27	31.04
DumPY : shorter than wild-type family member (dpy-2)	*Caenorhabditis elegans*	1-14	30.61
OSMotic avoidance abnormal family member (osm-10)	*Caenorhabditis elegans*	28-51	29.76
F46B6.10	*Caenorhabditis elegans*	6-89	29.15
DumPY : shorter than wild-type family member (dpy-10)	*Caenorhabditis elegans*	1-14	28.10
COLlagen family member (col-84)	*Caenorhabditis elegans*	1-14	28.10
F38A3.1	*Caenorhabditis elegans*	1-27	27.34
T10E10.1	*Caenorhabditis elegans*	1-27	27.33
COLlagen family member (col-2)	*Caenorhabditis elegans*	2-27	26.87

COLlagen family member (col-36)	*Caenorhabditis elegans*	1-27	26.73
F15H10.1	*Caenorhabditis elegans*	1-27	26.57
Nop10-like SnoRNP (144210) SCOP seed sequence: d2ey4e1		50-72	26.50
ROLler: helically twisted, animals roll when moving family member (rol-1)	*Caenorhabditis elegans*	1-14	26.24
C34F6.3	*Caenorhabditis elegans*	1-27	25.86
COLlagen family member (col-106)	*Caenorhabditis elegans*	2-27	25.77
COLlagen family member (col-166)	*Caenorhabditis elegans*	1-27	25.24
COLlagen family member (col-115)	*Caenorhabditis elegans*	1-14	24.41
CG13783-PA	*Drosophila melanogaster*	10-28	24.23
F15H10.2		1-27	24.06
Methylene tetrahydromethanopterin dehydrogenase		33-47	23.46
fixS protein	*Neisseria meningitidis*	6-19	23.35
F11G11.12	*Caenorhabditis elegans*	1-27	23.32
Chondrolectin precursor	*Homo sapiens*	2-20	23.15
F57B1.4	*Caenorhabditis elegans*	1-27	22.76
T21B4.2	*Caenorhabditis elegans*	1-27	22.47
Y69H2.14	*Caenorhabditis elegans*	1-27	22.39
F57B1.3	*Caenorhabditis elegans*	1-27	22.39
BLIstered cuticle family member (bli-2)	*Caenorhabditis elegans*	1-27	22.37
COLlagen family member (col-51)	*Caenorhabditis elegans*	2-27	22.30
ROLler: helically twisted, animals roll when moving family member (rol-8)	*Caenorhabditis elegans*	1-27	22.10
T10E10.6	*Caenorhabditis elegans*	1-27	22.08
DumPY : shorter than wild-type family member (dpy-10)	*Caenorhabditis elegans*	1-27	21.85
Virus attachment protein globular domain (49835) SCOP seed sequence: d1h7za_		50-68	21.78
Adenovirus fibre protein; cell receptor recognition, receptor	*Human adenovirus type 3*	44-68	21.71
COLlagen family member (col-165)	*Caenorhabditis elegans*	1-27	21.58
C44C10.1	*Caenorhabditis elegans*	1-27	21.26
Photosystem II reaction centre X protein (PsbX)	*Synechococcus sp. CC9311*	5-26	21.26
COLlagen family member (col-110)	*Caenorhabditis elegans*	2-27	20.99
DumPY : shorter than wild-type family member (dpy-9)	*Caenorhabditis elegans*	2-27	20.99

DumPY : shorter than wild-type family member (dpy-3)	*Caenorhabditis elegans*	1-27	20.77
COLlagen family member (col-34)	*Caenorhabditis elegans*	2-27	20.65
FAD/NAD-linked reductases, dimerisation (C-terminal) domain (55424) SCOP seed sequence: d1d7ya3		10-26	20.64
F17C11.3	*Caenorhabditis elegans*	2-27	20.63
COLlagen family member (col-173)	*Caenorhabditis elegans*	1-27	20.56
Secreted protein	*Streptomyces coelicolor*	4-19	20.37
K08F4.5	*Caenorhabditis elegans*	2-21	20.33
C34F6.2	*Caenorhabditis elegans*	2-27	20.30
COLlagen family member (col-124)	*Caenorhabditis elegans*	2-27	20.27
F32G8.5	*Caenorhabditis elegans*	2-25	20.14
I-TASSER			
Sec-independent protein translocase protein TatB	*Escherichia coli*		1.13
Sts-2 protein	*Mus musculus*		0.501
Predict Protein			
Protein binding		1-2, 52, 62	
Mitochondrial membrae			
Atome2			
Preprotein translocase SecA subunit	*Thermus thermophilus*		72.89
Deoxyribonuclease I	*Bos taurus*		63.10
FLAP endonuclease-1 protein	*Methanocaldococcus jannaschii*		58.69
E3 ubiquitin-protein ligase UBR2	*Homo sapiens*		48.38
Potassium large conductance calcium-activated channel, subfamily M, beta member 2	*Homo sapiens*		45.89
S-locus pollen protein	*Brassica rapa*		43.02
Regulatory protein SIR4	*Saccharomyces cerevisiae*		41.26
mRNA 3'-end-processing protein RNA14	*Kluyveromyces lactis*		36.60
Proliferating cell nuclear antigen	*Homo sapiens*		35.73
Protein (adenovirus fibre)	*Homo sapiens*		33.59
Fiber protein	*Human adenovirus 37*		31.21
Fiber protein	*Human adenovirus 2*		30.90
Adenovirus type 5 fiber protein	*Human adenovirus 5*		30.46
Fiber protein	*Human adenovirus 41*		24.60
Transmembrane protein 173	*Homo sapiens*		23.18

Stimulator of interferon genes protein	*Homo sapiens*		19.47
Oncogene product P14TCL1	*Homo sapiens*		16.55
HMTCP-1	*Homo sapiens*		13.22

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXVII. *Hyridella menziesii* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		80-85	99.25
TIGR01167 LPXTG cell wall anchor domain		8-35	99.10
TIGR03304 outer membrane insertion C-terminal signal		1-6	99.05
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		8-12	97.66
TIGR03501 GlyGly-CTERM domain		23-35	96.93
TIGR00756 pentatricopeptide repeat domain		60-69	93.52
MotB_plug Membrane MotB of proton-channel complex MotA/MotB.		17-38	87.38
Flagellar motor protein	*Bacillus subtilis*	1-38	87.21
CD274 antigen	*Homo sapiens*	2-65	86.19
Partially redundant sensor-transducer of the stress-activated PKC1-MPK1 signaling pathway	*Saccharomyces cerevisiae*	6-37	85.09
motB flagellar motor protein MotB		17-38	84.85, 84.34
MotB_plug: Membrane MotB of proton-channel complex MotA/MotB		17-38	84.08
motB flagellar motor protein MotB		1-38	82.76
Flagellar motor protein MotS		5-38	82.25
Flagellar motor protein MotS		17-38	80.86
Glycophorin		17-38	80.00

Flagellar motor protein MotD		1-38	79.60
Flagellar motor protein MotB	*Escherichia coli*	3-38	79.38
Basigin	*Mus musculus*	3-73	77.67
MEGF11 protein	*Homo sapiens*	8-37	77.66
Cell division protein	*Yersinia pestis*	6-40	77.15
Transmembrane glycoprotein A33 precursor	*Homo sapiens*	1-36	76.89
Flagellar motor protein	*Yersinia pestis*	5-38	76.41
Leukocyte-associated immunoglobulin-like receptor 1 isoform b precursor	*Homo sapiens*	5-43	76.34
Flagellar motor protein	*Bacillus subtilis*	17-38	76.34
C35D10.1	*Caenorhabditis elegans*	14-42	76.14
EGF-like-domain, multiple 9	*Homo sapiens*	2-37	76.07
EGF-like-domain, multiple 9	*Homo sapiens*	2-37	76.07
Carbamoyl-phosphate synthase L chain, ATP-binding	*Nostoc punctiforme*	7-46	76.06
MEGF10 protein	*Homo sapiens*	2-37	75.94
Golgi membrane protein with similarity to mammalian CASP	*Saccharomyces cerevisiae*	10-38	75.88
motB flagellar motor protein MotB		17-38	75.30
RIKEN cDNA 2900064A13	*Mus musculus*	14-39	75.22
CG18146-PB, isoform B	*Drosophila melanogaster*	17-37	74.16
Syntaxin 7	*Homo sapiens*	11-39	73.60
Glycoprotein A33 (transmembrane)	*Mus musculus*	1-36	73.22
CG31136-PA	*Drosophila melanogaster*	17-37	71.97
Chain length determinant protein	*Beggiatoa sp. PS*	8-38	71.77
Kin of IRRE-like 2	*Mus musculus*	18-86	71.74
Neuregulin 4	*Mus musculus*	3-36	71.40
motB flagellar motor protein MotB; Validated		17-38	71.22
Vesicle-associated membrane protein 1 isoform 1	*Homo sapiens*	17-38	71.16
STL2P	*Arabidopsis thaliana*	4-39	70.59
Flagellar motor protein	*Yersinia pestis*	17-39	70.30
RCR		18-38	68.98
NHL12	*Arabidopsis thaliana*	18-56	68.76
ZK353.4	*Caenorhabditis elegans*	12-34	68.51
Flagellar motor protein	*Pseudomonas aeruginosa*	17-38	67.85
VAMP-5_synaptobrevin		17-37	67.62
T20D4.12	*Caenorhabditis elegans*	17-48	67.04

CCAAT displacement protein isoform c	*Homo sapiens*	3-39	66.72
SIT: SHP2-interacting transmembrane adaptor protein, SIT		18-42	66.62
Endomucin		8-42	66.54
CCAAT displacement protein isoform b	*Homo sapiens*	3-39	66.34
Capsular polysaccharide biosynthesis protein Cap1A	*Staphylococcus aureus*	11-39	66.20
SYP61	*Arabidopsis thaliana*	11-36	65.62
ATP binding / kinase/ protein kinase/ protein serine/threonine kinase/ protein-tyrosine kinase	*Arabidopsis thaliana*	1-92	65.08
Regulator of length of O-antigen component of lipopolysaccharide chains	*Escherichia coli*	11-38	65.00
Essential cell division protein	*Escherichia coli*	6-40	64.98
YLS9	*Arabidopsis thaliana*	7-56	64.93
Integrin alpha-IIB	*Homo sapiens*	8-40	64.81
F11 receptor	*Mus musculus*	18-65	64.72
SIT SHP2-interacting transmembrane adaptor protein		18-42	64.07
Vesicle transport through interaction with t-SNAREs homolog 1A	*Mus musculus*	17-39	63.82
I-TASSER			
Type I restriction-modification system methyltransferase subunit	*Vibrio vulnificus*		0.527
Poly(ADP-ribose) glycohydrolase	*Rattus norvegicus*		0.512
Transporter	*Aquifex aeolicus*		0.505
Mre11 nuclease	*Pyrococcus furiosus*		0.503
Predict Protein			
Protein binding		1, 38-40, 54-56, 75-77, 79, 81	
Secreted			
Atome2			
Nucleoprotein	*Influenza A virus*		80.49
Tankyrase-1	*Mus musculus*		56.43
Carboxypeptidase A1	*Bos taurus*		45.65
Stromal cell-derived factor 1	*Homo sapiens*		42.76

Integrin alpha-IIb **(3)**	*Homo sapiens*		28.46-37.34
Na, K-ATPase alpha subunit	*Squalus acanthias*		33.92
Integrin alpha-1	*Homo sapiens*		33.59
HIG1 domain family member 1A	*Homo sapiens*		33.37
Sodium/potassium-transporting ATPase subunit alpha-1	*Sus scrofa*		32.09
Pulmonary surfactant-associated polypeptide C	*Sus scrofa*		31.12
Phospholemman	*Homo sapiens*		31.06
Importin subunit alpha-2	*Mus musculus*		30.90
T-cell surface glycoprotein CD4	*Homo sapiens*		30.38
SERCA1a	*Oryctolagus cuniculus*		29.56
Potassium channel protein RCK4	*Homo sapiens*		29.40
Vesicle-associated membrane protein 2	*Rattus norvegicus*		27.79
Beta-type platelet-derived growth factor receptor	*Homo sapiens*		27.49
Integrin alpha-IIb light chain	*Homo sapiens*		26.38

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXVIII. *Lasmigona complanata* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR01167 LPXTG cell wall anchor domain		54-59	99.46
TIGR04294 prepilin-type processing-associated H-X9-DG domain		18-21	99.32
TIGR03304 outer membrane insertion C-terminal signal		34-35	99.27
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		46-53	98.05
TIGR03501 GlyGly-CTERM domain		8-18	97.32
TIGR00756 pentatricopeptide repeat domain		38-45	94.58
CG7685-PA	*Drosophila melanogaster*	2-26	93.27
Intra-Golgi v-SNARE, required for transport of proteins between an early and a later Golgi compartment	*Saccharomyces cerevisiae*	3-26	84.46

TMEM156: TMEM156 protein family **(2)**		1-27	78.90, 78.70
Golgi SNARE BET1-related		3-24	71.61
CG13969-PA	*Drosophila melanogaster*	3-44	71.25
Ceramidase		3-44	70.74
Sensor histidine kinase	*Streptococcus pneumoniae*	1-40	69.88
LptF_YjgP LPS export ABC transporter permease LptF		1-29	67.65
Sensory box histidine kinase PhoR	*Staphylococcus aureus*	1-36	67.50
Saliv_gland_allergen_Aed3		2-19	64.97
Conserved inner membrane protein	*Escherichia coli*	1-29	63.38
T27F7.3a	*Caenorhabditis elegans*	2-36	62.64
GRP: Glycine rich protein family		6-26	61.27
Sterol reductase/lamin B receptor		19-47	60.88
Essential SNARE protein localized to the ER	*Saccharomyces cerevisiae*	5-26	60.56
Alkaline ceramidase 2	*Mus musculus*	3-44	58.05
Vesicle-associated membrane-associated protein		4-26	57.11
Human EMeRin homolog family member (emr-1)	*Caenorhabditis elegans*	6-23	56.93
GRP Glycine rich protein family		5-26	56.69
CbiN ABC-type cobalt transport system, periplasmic component		1-36	55.84
W02F12.2	*Caenorhabditis elegans*	3-53	55.64
Related to YPC1 - Alkaline ceramidase		3-44	55.42
Lipoprotein required for capsular polysaccharide translocation through the outer membrane	*Escherichia coli K12*	1-19	54.31
LptG_lptG LPS export ABC transporter permease LptG		1-29	53.94
CG11020-PA, isoform A	*Drosophila melanogaster*	1-38	52.93
CG3066-PD, isoform D	*Drosophila melanogaster*	2-27	52.93
SVM_signal: SVM protein signal sequence		2-23	51.26
DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	1-41	50.82
Alkaline ceramidase that also has reverse (CoA-independent) ceramide synthase activity	*Saccharomyces cerevisiae*	3-44	49.99
Retinoblastoma-associated protein	*Homo sapiens*	29-46	49.30
Protein transporter	*Arabidopsis thaliana*	3-26	48.30
Signal transduction histidine kinase	*Lactobacillus casei*	2-36	46.70
N-acylsphingosine amidohydrolase 3	*Homo sapiens*	3-44	45.65
F59E11.5	*Caenorhabditis elegans*	2-31	44.97

Ceramidase		3-44	44.88
Cytoplasmic membrane protein	*Bartonella henselae*	1-34	44.46
SYP125; t-SNARE	*Arabidopsis thaliana*	6-29	44.28
Rhodanese-like protein	*Beggiatoa sp. PS*	2-22	43.79
MORN repeat protein	*Beggiatoa sp. PS*	1-21	43.59
Protein containing DUF1239	*Beggiatoa sp. PS*	1-22	43.55
Y41D4B.24	*Caenorhabditis elegans*	3-34	43.31
Y110A7A.11	*Caenorhabditis elegans*	5-26	43.12
COLlagen family member (col-102)	*Caenorhabditis elegans*	1-46	42.45
C46H11.8	*Caenorhabditis elegans*	6-20	42.31
Vesicle-associated membrane protein	*Mus musculus*	4-26	42.29
SrtB		1-34	42.20
Urinary protein (RUP)/acrosomal protein SP-10		1-27	41.83
ATCDS1; phosphatidate cytidylyltransferase	*Arabidopsis thaliana*	45-77	41.81
Temporarily Assigned Gene name family member (tag-254)	*Caenorhabditis elegans*	6-21	41.51
Golgi phosphoprotein 2	*Homo sapiens*	1-39	41.51
Golgi phosphoprotein 2	*Homo sapiens*	1-39	41.51
RCR		8-23	41.47
BLASTP			
Membrane protein	*Enterococcus faecium*	2-77	1.00e-06
Membrane protein **(3)**	*Enterococcus faecium*	4-77	5.00e-06 – 1.00e-05
MULTISPECIES: membrane protein	*Enterococcus*	4-77	5.00e-06
Glycyl-tRNA synthetase subunit alpha	*Avibacterium paragallinarum*	23-77	2.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio litoralis*	23-77	2.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio rumoiensis*	23-77	2.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio mytili*	23-77	2.00e-05
COG0752 Glycyl-tRNA synthetase, alpha subunit	*uncultured bacterium B3TF_MPn_8*	23-77	3.00e-05
MULTISPECIES: glycyl-tRNA synthetase subunit alpha **(5)**	*Vibrio*	23-77	3.00e-05, 4.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio caribbeanicus*	23-77	3.00e-05
Glycyl-tRNA synthetase, partial	*Vibrio campbellii*	23-77	4.00e-05
Glycyl-tRNA synthetase alpha chain	*Vibrio sp. JCM 19241*	23-77	4.00e-05-

			1.00e-04
Glycyl-tRNA synthetase subunit alpha, partial **(2)**	*Vibrio parahaemolyticus*	23-77	4.00e-05
Glycyl-tRNA synthetase subunit alpha, partial	*Vibrio sp. ER1A*	23-77	4.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio shilonii*	23-77	4.00e-05
Glycine--tRNA ligase alpha subunit	*Vibrio splendidus*	23-77	4.00e-05
Glycyl-tRNA synthetase **(8)**	*Vibrio parahaemolyticus*	23-77	4.00e-05-1.00e-04
Glycyl-tRNA synthetase	*Vibrio campbellii*	23-77	4.00e-05
Glycyl-tRNA synthetase	*Vibrio sp. 090810a*	23-77	4.00e-05
Glycyl-tRNA synthetase	*Vibrio rotiferianus*	23-77	4.00e-05
Glycine--tRNA ligase alpha subunit	*Vibrio sagamiensis*	23-77	4.00e-05
MULTISPECIES: glycyl-tRNA synthetase subunit alpha	*Vibrio harveyi group*	23-77	4.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio tubiashii*	23-77	4.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio coralliilyticus*	23-77	4.00e-05
Glycyl-tRNA synthetase alpha chain	*Vibrio sp. C7*	23-77	4.00e-05
Glycyl-tRNA synthetase, partial	*Vibrio nigripulchritudo*	23-77	5.00e-05
Glycyl-tRNA synthetase alpha chain	*Vibrio ponticus*	23-77	6.00e-05
Glycyl-tRNA synthetase subunit alpha	*Vibrio shilonii*	23-77	6.00e-05
Glycyl-tRNA synthetase alpha chain	*Vibrio variabilis*	23-77	6.00e-05
Deacylase	*Maribacter sp. HTCC2170*	3-53	7.00e-05
Glycyl-tRNA synthetase alpha chain	*Vibrio sp. JCM 19236*	23-77	7.00e-05
Glycyl-tRNA synthetase alpha chain	*Vibrio sp. JCM 19231*	23-77	2.00e-04
Lebocin-like antibacterial protein	*Heliothis virescens*	6-76	2.00e-04
2-oxoglutarate dehydrogenase E2	*Staphylococcus hominis*	34-77	4.00e-04
P2Y purinoceptor 1, partial	*Podiceps cristatus*	1-49	5.00e-04
Transporter	*Rickettsia typhi*	8-44	5.00e-04
Transporter	*Rickettsia prowazekii*	8-44	6.00e-04
P2Y purinoceptor 1, partial	*Gavia stellata*	1-49	6.00e-04
Chemotaxis protein	*Lactobacillus parafarraginis*	15-74	6.00e-04
Permease	*Rickettsia prowazekii*	28-65	7.00e-04
I-TASSER			
Fimbrial protein (Pilin)	*Peptoclostridium difficile*		0.667
Residues 29-152, plus four N-terminal residues from the expression construct	*Neisseria meningitidis*		0.622
Wnt inhibitor of Dorsal protein (N-terminal domain-linker)	*Drosophila melanogaster*		0.618
Cytochrome P450ERYF	*Saccharopolyspora*		0.612

	erythraea		
Cytochrome P450 cypX	*Bacillus subtilis*		0.607
Inositol-1-monophosphatase	*Mycobacterium tuberculosis*		0.606
Cytochrome P450 119	*Sulfolobus solfataricus*		0.605
Cytochrome P450 107B1 (P450CVIIB1)	*Streptomyces himastatinicus*		0.604
Oxy protein	*Actinoplanes teichomyceticus*		0.602
367aa long hypothetical cytochrome P450	*Sulfolobus tokodaii*		0.600
Predict Protein			
Protein binding		1-2, 5, 31-32, 34-35, 48, 51, 57	
Mitochondrial membrane			
Atome2			
P fimbrial regulatory protein KS71A	*Escherichia coli*		92.97
Protein (neamatode anticoagulant protein C2)	*Ancylostoma caninum*		62.39
Herpes simplex virus protein ICP47 (active domain)	*Herpes simplex virus*		46.61
Polyribonucleotide nucleotidyltransferase	*Escherichia coli*		42.79
Neurotoxin BmP03	*Mesobuthus martensii*		41.75
Calcium-gated potassium channel mthK	*Methanothermobacter thermautotrophicus*		38.42
CREB-binding protein	*Mus musculus*		37.95
Polyribonucleotide nucleotidyltransferase	*Escherichia coli*		34.92
Protein translocase subunit secA	*Thermotoga maritima*		33.09
Protein translocase subunit secA	*Bacillus subtilis*		33.06
Protein-export membrane protein secG	*Escherichia coli*		31.50
CPAP	*Danio rerio*		26.35
Transcription factor Dp-1	*Homo sapiens*		24.47
FAB	*Mus musculus*		16.70

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of

10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXIX. *Toxolasma lividus* F-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		34-36	99.23
TIGR03304 outer membrane insertion C-terminal signal		1-8	99.14
TIGR01167 LPXTG cell wall anchor domain		95-96	98.81
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		18-22	97.99
GlyGly-CTERM domain		50-60	97.08
TIGR00756 pentatricopeptide repeat domain		30-49	94.54
ComGC		45-62	91.25
Stage III sporulation protein AF		18-63	80.41
Protein-export membrane protein		34-111	75.33
Transducer protein Htr37		24-64	74.74
Syntaxin-like t-SNARE		37-98	70.94
Mutants block sporulation after engulfment (stage III sporulation)		18-63	70.72
VAMP-5_synaptobrevin		36-68	67.81
CG11815-PA	*Drosophila melanogaster*	77-99	66.44
C-type LECtin family member (clec-35)	*Caenorhabditis elegans*	28-110	65.05
ComGC Competence protein ComGC		45-62	62.86
COLlagen family member (col-14)	*Caenorhabditis elegans*	21-65	58.40
Pili subunits (54523) SCOP seed sequence: d2pila_		4-62	56.94
General secretion pathway protein H	*Nostoc punctiforme*	28-63	56.08
Methyl-accepting chemotaxis protein	*Beggiatoa sp. PS*	33-59	55.50
Stage III sporulation protein AF (Spore_III_AF)		3-63	54.93
Opacity-associated protein A N-terminal motif		40-61	53.84
Protein involved in cis-Golgi membrane traffic; v-SNARE	*Saccharomyces cerevisiae*	25-64	53.77
Pili subunits (54523) SCOP seed sequence: d1oqwa_		44-62	53.25
F08F8.8	*Caenorhabditis elegans*	19-64	53.15
DevC protein	*Nostoc punctiforme*	24-66	52.26

Transducer protein Htr36	*Haloferax volcanii*	26-64	50.87
Opacity-associated protein A N-terminal motif		40-61	50.30
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-34)	*Caenorhabditis elegans*	59-104	50.27
Flagellar M-ring protein	*Bacillus subtilis*	25-79	49.97
SecD-TM1 SecD export protein N-terminal TM region.		35-63	48.71
C-type LECtin family member (clec-25)	*Caenorhabditis elegans*	43-107	48.40
Vesicle transport through interaction with t-SNAREs 1B	*Homo sapiens*	25-64	48.33
Vesicle transport through interaction with t-SNAREs 1B homolog	*Mus musculus*	25-64	48.30
Y57G11C.4	*Caenorhabditis elegans*	25-63	47.57
C05E11.1	*Caenorhabditis elegans*	8-59	46.93
Methyl-accepting chemotaxis protein II	*Yersinia pestis*	31-64	46.49
Stage III sporulation protein AF (Spore_III_AF)		30-63	46.41
DevC protein	*Nostoc punctiforme*	26-65	45.60
DevC protein	*Nostoc punctiforme*	24-65	45.36
T10E10.5	*Caenorhabditis elegans*	25-66	44.80
CG3279-PA	*Drosophila melanogaster*	25-76	44.73
SYP123; t-SNARE	*Arabidopsis thaliana*	25-66	43.93
VTI12; SNARE binding/receptor	*Arabidopsis thaliana*	25-64	43.41
Protein export protein SecD	*Neisseria meningitidis*	35-67	43.40
DevC protein	*Nostoc punctiforme*	28-65	43.13
DevC protein	*Nostoc punctiforme*	24-65	42.27
VTI11; receptor	*Arabidopsis thaliana*	25-64	42.17
Vesicle-associated membrane protein 5 (myobrevin)	*Homo sapiens*	36-69	41.51
T24B1.1	*Caenorhabditis elegans*	31-69	41.28
Vesicle transport v-snare protein	*Schizosaccharomyces pombe*	25-64	41.03
Intra-Golgi v-SNARE, required for transport of proteins between an early and a later Golgi compartment	*Saccharomyces cerevisiae*	25-63	40.76
F41F3.3	*Caenorhabditis elegans*	42-61	40.59
TonB family protein	*Nostoc punctiforme 73102*	37-81	40.54
Related to VTI1 - v-SNARE: involved in Golgi retrograde protein traffic		25-64	39.71
Proline-rich region	*Synechococcus sp. CC9311*	23-66	39.54

Resistance to inhibitors of cholinesterase 3 homolog	*Homo sapiens*	41-105	39.30
F46F5.7	*Caenorhabditis elegans*	36-111	39.23
Protein export protein SecD	*Pseudomonas aeruginosa*	35-67	38.85
Methyl-accepting chemotaxis protein III	*Escherichia coli*	33-59	38.71
Competence protein CglC	*Streptococcus pneumoniae*	26-63	38.44
COLlagen family member (col-77)	*Caenorhabditis elegans*	25-66	37.87
v-SNARE (vesicle specific SNAP receptor)	*Saccharomyces cerevisiae*	19-64	37.76
Transcriptional accessory factor Tex **(2)**	*Pseudomonas aeruginosa*	79-99	37.60
F0F1 ATP synthase subunit A	*Mycobacterium tuberculosis*	18-67	37.34
a disintegrin and metalloproteinase domain 7	*Homo sapiens*	45-108	37.11
CG13581-PA	*Drosophila melanogaster*	101-113	36.38
Protein export protein SecD	*Escherichia coli*	35-67	36.37
SYP124; t-SNARE	*Arabidopsis thaliana*	25-64	36.06
Laeverin	*Homo sapiens*	40-111	35.41
COLlagen family member (col-174)	*Caenorhabditis elegans*	25-66	35.28
CG11500-PA	*Drosophila melanogaster*	5-89	35.23
SecD-TM1: SecD export protein N-terminal TM region		36-67	34.96
Multi-sensor signal transduction histidine kinase	*Nostoc punctiforme*	21-64	34.75
MacB_PCD MacB-like periplasmic core domain.		32-66	34.43
Type IV Pilin Pak	*Pseudomonas aeruginosa*	44-62	34.26
I-TASSER			
UNC-45 protein, SD10334p	*Drosophila melanogaster*		0.512
RCD1 required for cell differentiation1 homolog	*Homo sapiens*		0.509
Chloride intracellular channel exc-4	*Caenorhabditis elegans*		0.504
Telomerase-binding protein EST1A (tetratricopeptide repeat, residues 580-1166)			0.502
Protein UNC-45	*Caenorhabditis elegans*		0.502
Karyopherin alpha (armadillo domain)	*Saccharomyces cerevisiae*		0.500
Predict Protein			
Protein binding		1, 3, 5, 10-11, 14-18, 20-21, 31, 35, 38-39,	

		64-66, 68, 70, 72, 90, 94-95, 97	
Polynucleotide binding		27	
Mitochondrial membrane			
Atome2			
FLT3 ligand (receptor binding domain)	*Homo sapiens*		99.28
Protein parD	*Escherichia coli*		59.11
Intrinsic membrane protein pufX	*Rhodobacter sphaeroides*		54.16
Envelope protein E	*Dengue virus*		46.45
Neopetrosiamide A	*Neopetrosia sp.*		38.08
Laccase	*Rigidoporus microporus*		30.09
Laccase	*Botrytis aclada*		25.58
Ascorbate oxidase	*Cucurbita pepo*		23.65
Iron transport multicopper oxidase FET3	*Saccharomyces cerevisiae*		23.49
Laccase 1	*Coprinopsis cinerea*		22.43
Laccase	*Steccherinum ochraceum*		22.40
Fimbrial protein	*Neisseria gonorrhoeae*		20.56
Laccase-1	*Melanocarpus albomyces*		15.28

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXX. *Margaritifera margaritifera* F-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		23-24	99.14
TIGR04294 prepilin-type processing-associated H-X9-DG domain		44-49	99.11
TIGR01167 LPXTG cell wall anchor domain		32-48	98.87

TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		49-56	97.75
TIGR03501 GlyGly-CTERM domain		36-48	97.15
TIGR00756 pentatricopeptide repeat domain		16-23	93.83
T24B1.1	*Caenorhabditis elegans*	21-52	84.13
Occlusion-derived virus envelope protein ODV-E18		21-62	72.05
d.24.1 Pili subunits (54523) SCOP seed sequence: d2pila_		31-50	68.15
d.24.1 Pili subunits (54523) SCOP seed sequence: d1oqwa_		31-50	67.59
RCR		32-50	65.14
Cytochrome c550	*Bacillus subtilis*	27-61	62.79
Occlusion-derived virus envelope protein ODV-E18		23-55	62.79
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-90)	*Caenorhabditis elegans*	34-53	61.72
General secretion pathway protein H	*Beggiatoa sp. PS*	31-50	60.83
CytB-hydogenase Ni/Fe-hydrogenase, b-type cytochrome subunit		8-49	58.96
Activated in Blocked Unfolded protein response family member (abu-1)	*Caenorhabditis elegans*	34-53	58.69
Alpha defensin		39-50	58.16
ComB		6-50	57.29
COLlagen family member (col-34)	*Caenorhabditis elegans*	19-56	55.90
Secreted protein	*Beggiatoa sp. PS*	31-50	55.82
COLlagen family member (col-93)	*Caenorhabditis elegans*	19-56	54.87
Serine protease inhibitor		37-90	51.90
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-78)	*Caenorhabditis elegans*	34-55	49.09
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-2)	*Caenorhabditis elegans*	34-55	48.47
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-76)	*Caenorhabditis elegans*	34-55	48.10
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-79)	*Caenorhabditis elegans*	34-55	48.10
Integral membrane protein	*Streptomyces coelicolor*	1-55	47.76
W06F12.2a	*Caenorhabditis elegans*	10-52	47.72
C17H11.6c	*Caenorhabditis elegans*	8-59	47.62

Methyl-CpG BinDing protein family member (mbd-2)	*Caenorhabditis elegans*	9-21	47.44
RCR		32-50	46.40
COLlagen family member (col-91)	*Caenorhabditis elegans*	31-56	45.97
Methyl-CpG binding domain protein 3-like 1	*Mus musculus*	9-27	44.23
Pleiotrophin family member		34-53	43.89
F27E5.3	*Caenorhabditis elegans*	31-50	43.81
F420-nonreducing hydrogenase II subunit cytochrome B	*Methanosarcina mazei*	22-50	43.78
Crumbs homolog 2	*Homo sapiens*	23-53	43.61
General secretion pathway protein J	*Yersinia pestis*	31-50	43.53
F26B1.1	*Caenorhabditis elegans*	1-49	43.50
Glycine rich protein family		35-53	43.19
COLlagen family member (col-94)	*Caenorhabditis elegans*	19-56	43.14
Activated in Blocked Unfolded protein response family member (abu-7)	*Caenorhabditis elegans*	34-55	42.68
TetraSPanin family member (tsp-14)	*Caenorhabditis elegans*	33-109	42.50
K08F4.5	*Caenorhabditis elegans*	31-50	42.44
COLlagen family member (col-92)	*Caenorhabditis elegans*	19-56	42.30
Type II secretion system protein I.		31-52	42.13
COLlagen family member (col-139)	*Caenorhabditis elegans*	19-56	42.09
COLlagen family member (col-108)	*Caenorhabditis elegans*	19-56	41.97
F46H6.5	*Caenorhabditis elegans*	75-94	40.38
COLlagen family member (col-102)	*Caenorhabditis elegans*	31-56	40.22
CG2040-PA, isoform A	*Drosophila melanogaster*	19-45	40.06
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-91)	*Caenorhabditis elegans*	34-55	39.77
Type 4 fimbrial biogenesis protein FimT	*Pseudomonas aeruginosa PAO1*	31-50	39.48
T-cell receptor-associated transmembrane adapter 1		31-48	39.41
I-TASSER			
Tropomyosin	*Oryctolagus cuniculus*		1.03
Oligopeptidase	*Geobacillus sp. MO-1*		0.532
Glucose-6-phosphate isomerase	*Brucella melitensis*		0.523
Glucose-6-phosphate isomerase	*Vibrio cholerae*		0.518
Glucose-6-phosphate isomerase	*Plasmodium falciparum*		0.518
Glucose-6-phosphate isomerase	*Sus scrofa*		0.518
Cytochrome P450 107B1 (P450CVIIB1)	*Streptomyces*		0.517

	himastatinicus		
Glucose-6-phosphate isomerase	*Escherichia coli*		0.515
Oligoendopeptidase F	*Geobacillus stearothermophilus*		0.515
Phosphoglucose isomerase	*Geobacillus stearothermophilus*		0.512
Predict Protein			
Protein binding		1, 3-4, 16, 18-20, 24-29, 55, 57, 81-82	
Secreted			
Atome2			
Protein MXIG	*Shigella flexner*		86.27
Protein parD	*Escherichia coli*		76.28
ARF GTPase-activating protein GIT1	*Rattus norvegicus*		66.40
NifU-like protein, mitochondrial	*Saccharomyces cerevisiae*		43.28
Photosynthetic reaction center C subunit	*Thermochromatium tepidum*		39.97
Lichenicidin VK21 A1	*Bacillus licheniformis*		38.81
Collagen alpha 1 (heparin binding site)	*Gallus gallus*		38.00
Adenovirus fibre	*Human adenovirus 2*		27.29
Formate dehydrogenase, nitrate-inducible, major subunit	*Escherichia coli*		26.46
Fimbrial protein	*Dichelobacter nodosus*		22.77
Fimbrial protein	*Neisseria gonorrhoeae*		22.40
Fimbrial protein	*Pseudomonas aeruginosa*		20.49
Protein (LCA)	*Homo sapiens*		18.92
Fiber protein 2	*Human adenovirus 41*		18.06

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXI. *Anodonta anatina* F-*ORF* function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR01167 LPXTG cell wall anchor domain		55-60	99.47
TIGR04294 prepilin-type processing-associated H-X9-DG domain		19-22	99.31
TIGR03304 outer membrane insertion C-terminal signal		35-36	99.24
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		47-54	98.03
TIGR03501 GlyGly-CTERM domain		9-19	97.33
TIGR00756 pentatricopeptide repeat domain		26-46	94.32
CG7685-PA	*Drosophila melanogaster*	4-27	91.69
Intra-Golgi v-SNARE	*Saccharomyces cerevisiae*	2-27	86.02
CG13969-PA	*Drosophila melanogaster*	2-45	80.31
TMEM156 protein family		2-28	77.90
Ceramidase		3-45	74.10
Peptidoglycan-associated lipoprotein Pal	*Yersinia pestis*	1-20	69.76
Saliv_gland_allergen_Aed3		3-20	69.53
Retinoblastoma-associated protein	*Homo sapiens*	30-47	68.54
CG3066-PD, isoform D	*Drosophila melanogaster*	3-28	68.32
Alkaline ceramidase 2	*Mus musculus*	2-45	66.07
Golgi SNARE BET1-related		4-25	65.91
W02F12.2	*Caenorhabditis elegans*	2-54	65.41
ABC transporter, periplasmic amino acid-binding protein	*Bartonella henselae*	1-23	63.51
Undecaprenyl pyrophosphate phosphatase	*Escherichia coli*	1-28	62.85
Peptidoglycan-associated outer membrane lipoprotein	*Escherichia coli*	1-20	62.73
T27F7.3a	*Caenorhabditis elegans*	3-37	62.11
Cytochrome C-type biogenesis protein CcmE	*Pseudomonas aeruginosa*	1-43	61.79
Sterol reductase/lamin B receptor		20-48	61.33
Glycine rich protein family		6-27	61.32
SrtB		1-35	61.18
Human EMeRin homolog family member (emr-1)	*Caenorhabditis elegans*	6-24	61.00
Related to YPC1 - Alkaline ceramidase		2-45	60.98
Periplasmic heme chaperone	*Escherichia coli*	1-43	60.72
Protein transporter	*Arabidopsis thaliana*	2-27	59.33
Signal transduction histidine kinase	*Lactobacillus casei*	3-37	59.20
Essential SNARE protein localized to the ER	*Saccharomyces cerevisiae*	6-27	58.72

LPS export ABC transporter permease LptF		4-30	58.57
Syntaxin 5	*Mus musculus*	2-24	58.16
Vesicle-associated membrane protein-associated protein		5-27	57.33
SYP31; t-SNARE	*Arabidopsis thaliana*	2-24	57.00
N-acylsphingosine amidohydrolase 3	*Homo sapiens*	2-45	55.58
GRP Glycine rich protein family		6-27	55.38
Target membrane receptor (t-SNARE)	*Saccharomyces cerevisiae*	2-24	55.04
Alkaline ceramidase that also has reverse (CoA-independent) ceramide synthase activity	*Saccharomyces cerevisiae*	2-45	54.77
Soluble secreted antigen MPT53 precursor	*Mycobacterium tuberculosis*	1-28	54.23
Temporarily Assigned Gene name family member (tag-254)	*Caenorhabditis elegans*	7-22	53.72
Conserved inner membrane protein	*Escherichia coli*	4-30	53.59
Ceramidase		2-45	52.04
F59E11.5	*Caenorhabditis elegans*	3-32	51.37
Cytochrome C-type protein NapC	*Beggiatoa sp. PS*	2-36	50.69
Syntaxin 7	*Homo sapiens*	2-24	50.07
N-acylsphingosine amidohydrolase 3-like	*Homo sapiens*	2-45	49.95
Syntaxin-related protein required for vacuolar assembly	*Saccharomyces cerevisiae*	2-24	49.79
Y59E9AL.7	*Caenorhabditis elegans*	2-25	49.43
Rhodanese-like protein	*Beggiatoa sp. PS*	3-28	49.25
PAP2 family protein	*Staphylococcus aureus*	1-35	48.90
Secreted protein	*Beggiatoa sp. PS*	1-36	48.75
Peptidoglycan associated lipoprotein OprL precursor	*Pseudomonas aeruginosa PAO1*	1-20	48.67
Sortase B	*Staphylococcus aureus subsp. aureus COL*	1-42	48.38
SYP125; t-SNARE	*Arabidopsis thaliana*	7-30	47.94
Y41D4B.24	*Caenorhabditis elegans*	2-35	47.44
CG14084-PB, isoform B	*Drosophila melanogaster*	2-25	47.44
I-TASSER			
Glucokinase regulatory protein	*Homo sapiens*		0.522
Glucokinase regulatory protein	*Xenopus laevis*		0.517
Cation exchanger YfkE	*Bacillus subtilis*		0.512

Pathogenicity island 1 effector protein	*Chromobacterium violaceum*		0.509
Unconventional myosin-Va	*Mus musculus*		0.508
Methane monooxygenase hydroxylase	*Methylosinus trichosporium*		0.505
Inositol-1-monophosphatase	*Mycobacterium tuberculosis*		0.505
Predict Protein			
Protein binding		1-3, 6, 32-33, 56-58	
Mitochondrial membrane			
Atome2			
Transposon Tn557 toxic shock syndrome toxin-1	*Staphylococcus aureus*		78.60
S67	*Sicarius dolichocephalus*		66.70
BirA bifunctional protein	*Escherichia coli*		61.21
SERCA1a	*Oryctolagus cuniculus*		55.57
Sodium/potassium-transporting ATPase subunit alpha-1	*Sus scrofa*		51.63
Antitoxin RelB3	*Methanocaldococcus jannaschii*		48.98
Na, K-ATPase alpha subunit	*Squalus acanthias*		46.29
Transcription factor Dp-1			45.74
Nicotinic acetylcholine receptor	*Torpedo californica*		41.21
CPAP	*Danio rerio*		32.69
Potassium-transporting ATPase alpha	*Sus scrofa*		32.06
Vesicle-associated membrane protein 2	*Rattus norvegicus*		8.39

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXII. *Utterbackia imbecillis* H-*ORF* sequence 1 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		201-204	99.30
TIGR03304 outer membrane insertion C-terminal signal		49-52	99.27
TIGR01167 LPXTG cell wall anchor domain		75-77	98.89
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		21-23	97.49
TIGR03501 GlyGly-CTERM domain		49-57	96.92
CG12522-PA	*Drosophila melanogaster*	77-147	97.70
CG12522-PA	*Drosophila melanogaster*	74-137	97.65
G protein-coupled receptor 152	*Homo sapiens*	1-156	96.83
G protein-coupled receptor 152	*Homo sapiens*	3-156	96.50
Procyclic acidic repetitive protein (PARP)		80-149	96.21
R06C7.4	*Caenorhabditis elegans*	75-155	96.17
Procyclic acidic repetitive protein (PARP)		84-153	95.96
R06C7.4	*Caenorhabditis elegans*	74-155	95.67
K09E4.6	*Caenorhabditis elegans*	79-156	95.13
TIGR00756 pentatricopeptide repeat domain		14-21	94.13
T14A8.2	*Caenorhabditis elegans*	34-155	95.01
Related to CSR1 - phosphatidylinositol transfer protein		72-153	94.46
K09E4.6	*Caenorhabditis elegans*	88-156	94.22
T06A4.1b	*Caenorhabditis elegans*	73-153	93.81
T06A4.1b	*Caenorhabditis elegans*	73-155	93.76
Armadillo repeat containing, X-linked 4	*Homo sapiens*	75-153	93.72
DumPY: shorter than wild-type family member (dpy-10)	*Caenorhabditis elegans*	1-66	93.60
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	74-156	93.46
Copper-binding protein	*Methanosarcina mazei*	11-154	93.45
Prolipoprotein diacylglyceryl transferase	*Mycobacterium tuberculosis*	21-156	93.44
Related to CSR1 - phosphatidylinositol transfer protein		75-156	92.94

T04F8.8	*Caenorhabditis elegans*	85-146	92.56
T04F8.8	*Caenorhabditis elegans*	75-136	92.47
F56B6.4a	*Caenorhabditis elegans*	74-152	92.43
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	75-153	92.40
Armadillo repeat containing, X-linked 4	*Homo sapiens*	73-157	92.39
F56B6.4a	*Caenorhabditis elegans*	74-156	92.37
Junctophilin 2	*Mus musculus*	74-212	92.12
SphingoMyelin Synthase family member (sms-1)	*Caenorhabditis elegans*	75-204	91.33
Junctophilin 1	*Mus musculus*	74-212	91.12
BLIstered cuticle family member (bli-2)	*Caenorhabditis elegans*	12-67	91.06
Solute carrier family 16, member 2	*Homo sapiens*	75-149	91.03
CG1468-PA	*Drosophila melanogaster*	71-156	90.90
Solute carrier family 16, member 2	*Homo sapiens*	74-155	90.73
CG12316-PA, isoform A	*Drosophila melanogaster*	84-150	90.67
CG12316-PB, isoform B	*Drosophila melanogaster*	84-150	90.67
Eukaryotic translation initiation factor 3, subunit 9	*Mus musculus*	76-156	90.65
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	74-157	90.55
Eukaryotic translation initiation factor 3, subunit 9	*Mus musculus*	69-156	90.48
Junctophilin 1	*Homo sapiens*	71-212	90.44
Adhesion exoprotein	*Lactobacillus casei*	72-156	89.55
F57B1.3	*Caenorhabditis elegans*	7-67	89.46
Membralin isoform 1	*Homo sapiens*	7-156	89.25
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	74-156	88.92
Diacylglycerol kinase kappa	*Homo sapiens*	73-156	88.91
F47B8.5	*Caenorhabditis elegans*	74-155	88.65
F47B8.5	*Caenorhabditis elegans*	75-156	88.53
T10E10.5	*Caenorhabditis elegans*	12-67	88.21
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	12-63	87.68
CG1468-PA	*Drosophila melanogaster*	74-156	87.29
DumPY: shorter than wild-type family member (dpy-14)	*Caenorhabditis elegans*	20-67	87.27
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	74-153	87.09
Diacylglycerol kinase kappa	*Homo sapiens*	72-156	86.90
V-set and immunoglobulin domain containing 1	*Mus musculus*	74-155	86.33
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	75-156	85.78
T10G3.1	*Caenorhabditis elegans*	73-154	85.56

T10G3.1	*Caenorhabditis elegans*	73-154	85.23
Adhesion exoprotein	*Lactobacillus casei*	76-156	84.94
R09F10.3	*Caenorhabditis elegans*	72-156	84.49
F33A8.9	*Caenorhabditis elegans*	10-67	84.46
COLlagen family member (col-110)	*Caenorhabditis elegans*	21-68	84.41
D1054.11	*Caenorhabditis elegans*	72-156	84.32
Y54E10BL.2	*Caenorhabditis elegans*	22-67	84.00
COLlagen family member (col-102)	*Caenorhabditis elegans*	25-67	83.99
PF70 protein	*Plasmodium falciparum*	74-154	83.94
C09F9.2	*Caenorhabditis elegans*	72-153	83.50
F11G11.11	*Caenorhabditis elegans*	16-67	83.21
BLASTP, PSIBLAST			
Bv80/Bb-1, partial	*Babesia bovis*	77-152	2e-10, 8e-14
Bv80/Bb-1, partial	*Babesia bovis*	76-157	4e-10, 2e-13
Bv80/Bb-1, partial	*Babesia bovis*	6-152	2e-09, 6e-13
S-layer protein precursor	*Bacillus thuringiensis*	73-152	6e-09, 2e-12
Cell surface protein, partial	*Bacillus thuringiensis*	73-152	7e-09, 3e-12
85 kDa protein	*Babesia bovis*	77-152	3e-08, 1e-11
Bv80/Bb-1, partial **(3)**	*Babesia bovis*	76-132	1e-07- 2e-10
Cell surface protein **(2)**	*Bacillus thuringiensis*	75-152	1e-07- 4e-11
Bv80, partial	*Babesia bovis*	76-136	1e-07, 5e-11
85 kDa protein **(2)**	*Babesia bovis*	76-182	1e-06-3e-10
LdORF-129 peptide	*Lymantria dispar multiple nucleopolyhedrovirus*	74-144	2e-06, 7e-10
ORF-132 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	74-131	4e-06, 2e-09

GH24581	*Drosophila grimshawi*	79-143	6e-05, 2e-08
Type I restriction modification protein	*Mycoplasma pneumoniae*	76-157	2e-04, 9e-08
Restriction endonuclease, S subunit	*Mycoplasma pneumoniae*	70-157	0.005, 2e-06
Type I restriction modification protein	*Mycoplasma pneumoniae*	70-157	0.005, 2e-06
Restriction endonuclease, S subunit	*Mycoplasma pneumoniae*	62-157	0.007, 3e-06
PSIBLAST			
Protein B602L, partial	*Columba livia*	76-154	6e-11
Bv80, partial	*Babesia bovis*	76-132	3e-08
ORF-126 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	72-140	4e-08
B602L, partial **(2)**	*African swine fever virus*	60-148	5e-08, 2e-07
Central variable region protein **(2)**	*African swine fever virus*	60-154	6e-08-7e-07
Central variable region protein	*African swine fever virus*	60-130	7e-08
pB602L	*African swine fever virus*	60-132	8e-08
Bv80, partial	*Babesia bovis*	76-152	8e-08
9RL protein	*African swine fever virus*	65-153	8e-08
B602L, partial	*African swine fever virus*	60-148	1e-07
9RL, partial **(3)**	*African swine fever virus*	60-129	1e-07-3e-05
Bv80, partial **(3)**	*Babesia bovis*	80-132	2e-07, 2e-06
Response regulator receiver domain protein (CheY-like)	*Nodularia spumigena*	70-136	2e-07
B602L protein	*African swine fever virus*	66-153	3e-07
9RL protein **(2)**	*African swine fever virus*	60-154	3e-07-9e-07
U1	*Hyposoter didymator ichnovirus*	77-138	3e-07
Pathway-specific nitrogen regulator	*Metarhizium guizhouense*	68-132	4e-07
9RL protein	*African swine fever virus*	65-136	4e-07

B602L protein, partial **(3)**	*African swine fever virus*	65-169	4e-07-6e-06
BV80 merozoite protein	*Babesia bovis*	76-152	5e-07
Bv80, partial	*Babesia bovis*	80-136	6e-07
Bv80/Bb-1, partial **(5)**	*Babesia bovis*	80-128	6e-07-2e-06
9RL protein, partial	*African swine fever virus*	65-137	7e-07
9RL protein, partial	*African swine fever virus*	75-147	7e-07
GG21615	*Drosophila erecta*	65-140	9e-07
B602L protein **(2)**	*African swine fever virus*	66-152	1e-06, 3e-04
Bv80/Bb-1, partial	*Babesia bovis*	78-128	1e-06
9RL protein	*African swine fever virus*	65-164	2e-06
Type IV secretion protein Rhs, partial	*Nocardioides sp. URHA0020*	76-155	2e-06
9RL protein	*African swine fever virus*	65-154	2e-06
9RL protein, partial	*African swine fever virus*	81-165	2e-06
B602L protein **(2)**	*African swine fever virus*	66-130	2e-06, 1e-04
B602L, partial	*African swine fever virus*	60-130	3e-06
9RL, partial	*African swine fever virus*	60-125	3e-06
B602L, partial **(4)**	*African swine fever virus*	60-144	4e-06-1e-05
pB602L, partial	*African swine fever virus*	65-145	5e-06
B602L protein	*African swine fever virus*	66-154	5e-06
9RL	*African swine fever virus*	75-164	6e-06
9RL protein **(2)**	*African swine fever virus*	65-130	6e-06, 1e-04
GL22603	*Drosophila persimilis*	81-142	7e-06
B602L, partial **(4)**	*African swine fever virus*	60-129	8e-06-5e-05
Translation initiation factor eIF2B	*Metarhizium robertsii*	68-129	8e-06
B602L protein **(2)**	*African swine fever virus*	65-152	9e-06, 2e-05
B602L protein	*African swine fever virus*	65-148	9e-06
Pathway-specific nitrogen regulator	*Metarhizium anisopliae*	68-129	9e-06
Pathway-specific nitrogen regulator, partial	*Metarhizium brunneum*	68-129	9e-06

B602L protein	*African swine fever virus*	75-164	1e-05
9RL protein, partial	*African swine fever virus*	81-147	1e-05
Mucin	*Trichomonas vaginalis*	71-140	1e-05
Cell surface protein **(2)**	*Bacillus thuringiensis*	23-152	1e-05
B602L, partial	*African swine fever virus*	60-169	1e-05
B602L protein **(2)**	*African swine fever virus*	70-130	1e-05, 4e-04
S-layer protein	*Bacillus thuringiensis*	23-152	1e-05
Outer membrane autotransporter barrel domain-containing protein	*Escherichia coli*	68-124	2e-05
Central variable region protein	*African swine fever virus*	60-134	2e-05
9RL protein	*African swine fever virus*	65-125	2e-05
B602L, partial	*African swine fever virus*	64-148	2e-05
9RL protein, partial	*African swine fever virus*	79-153	2e-05
B602L protein	*African swine fever virus*	66-153	2e-05
9RL protein **(2)**	*African swine fever virus*	75-130	2e-05, 0.003
B602L protein **(3)**	*African swine fever virus*	65-140	2e-05- 4e-05
BA71V-B602L	*African swine fever virus*	60-132	3e-05
9RL, partial	*African swine fever virus*	60-129	3e-05
Cellulosomal scaffoldin anchoring protein	*Trichomonas vaginalis*	76-152	3e-05
GE10809	*Drosophila yakuba*	74-148	3e-05
CG3108	*Drosophila melanogaster*	73-146	3e-05
B602L **(2)**	*African swine fever virus*	70-129	4e-05, 5e-05
9RL	*African swine fever virus*	81-140	4e-05
GE16785	*Drosophila yakuba*	84-137	4e-05
Ribonuclease E	*Nitrincola lacisaponensis*	77-146	4e-05
pB602L **(2)**	*African swine fever virus*	72-132	4e-05, 5e-05
B602L protein **(2)**	*African swine fever virus*	71-129	5e-05
B602L	*African swine fever virus*	60-152	5e-05
9RL protein **(2)**	*African swine fever virus*	65-129	5e-05, 6e-05
9RL protein **(2)**	*African swine fever virus*	65-140	6e-05

9RL **(3)**	*African swine fever virus*	75-148	6e-05-5e-04
B602L protein	*African swine fever virus*	75-140	7e-05
GI15252	*Drosophila mojavensis*	73-148	8e-05
Type I restriction modification protein, partial	*Mycoplasma pneumoniae*	96-157	8e-05
B602L	*African swine fever virus*	73-130	9e-05
B602L protein	*African swine fever virus*	81-140	1e-04
B602L protein **(8)**	*African swine fever virus*	75-148	1e-04-2e-04
Translation initiation factor IF-2	*Eubacterium sp. CAG:786*	82-153	1e-04
B602L protein	*African swine fever virus*	65-129	1e-04
Elicitin-like protein 6 precursor, partial	*Phytophthora medicaginis*	74-150	1e-04
9RL	*African swine fever virus*	74-140	2e-04
9RL	*African swine fever virus*	74-144	2e-04
B602L protein	*African swine fever virus*	76-148	2e-04
Rogdi domain containing protein	*Haemonchus contortus*	80-138	2e-04
9RL	*African swine fever virus*	75-144	4e-04
B602L protein	*African swine fever virus*	66-148	5e-04
Transcription factor IIIB 50 kDa subunit	*Xenopus tropicalis*	76-130	5e-04
Involucrin repeat protein	*Ophiostoma piceae*	74-136	6e-04
Peptidase	*Actinoplanes sp. SE50/110*	77-127	6e-04
B602L protein	*African swine fever virus*	75-130	0.001
Prokaryotic cytochrome b561 family protein	*Burkholderia pseudomallei*	76-127	0.001
B-type cytochrome	*Burkholderia pseudomallei*	76-127	0.001
9RL	*African swine fever virus*	74-130	0.001
Thylakoid rhodanese-like protein	*Medicago truncatula*	73-132	0.001
9RL	*African swine fever virus*	74-129	0.002
Cell division protein FtsK	*Carnobacterium sp. 17-4*	79-179	0.003
Autotransporter protein, partial	*Escherichia coli*	73-122	0.003
I-TASSER			
Survival motor neuron protein **(3)**	*Homo sapiens*		1.14-1.81
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		2.21
Myc box dependent interacting protein 1**(2)**	*Homo sapiens*		1.00-1.90
Accumulation associated protein **(3)**	*Staphylococcus epidermidis*		1.13-1.97
HIV-1 capsid	*Human immunodeficiency virus 1*		0.513
Gag Polyprotein	*Human immunodeficiency virus 1*		0.510

Capsid protein P24	*Human immunodeficiency virus 2*		0.504
Predict Protein			
Protein binding		1, 221-222	
Secreted			
Diacylglycerol kinase kappa **(3)**	*Homo sapiens*		5e-04-0.002
Proteoglycan 4 **(2)**	*Homo sapiens*		0.17, 0.99
Proteoglycan 4 **(9)**	*Mus musculus*		1e-08-0.002
Atome2			
Lamin A/C	*Homo sapiens*		92.75
Protein bicaudal D	*Drosophila melanogaster*		84.72
80 kDa MCM3-associated protein	*Homo sapiens*		76.71
Selenoprotein S	*Homo sapiens*		72.30
Heat shock protein	*Saccharomyces cerevisiae*		50.07
Nucleoprotein	*Andes virus*		44.18
Herpes simplex virus protein ICP47	*Herpes simplex virus 1*		37.48
Cupiennin-1a	*Cupiennius salei*		35.85
RNA-binding protein 5	*Homo sapiens*		29.47

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXIII. *Utterbackia imbecillis* H-*ORF* sequence 2 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		45-50	99.28
TIGR03304 outer membrane insertion C-terminal		75-81	99.22

signal			
TIGR01167 LPXTG cell wall anchor domain		56-72	98.94
G protein-coupled receptor 152	*Homo sapiens*	2-227	98.54
G protein-coupled receptor 152	*Homo sapiens*	4-229	98.36
Y51F10.4a	*Caenorhabditis elegans*	6-229	98.32
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		227-231	97.63
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	46-218	98.28
Integral membrane protein	*Streptomyces coelicolor*	44-224	98.11
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	51-222	98.01
CG17010-PA, isoform A	*Drosophila melanogaster*	103-227	97.89
TIGR03501 GlyGly-CTERM domain		58-71	97.07
Integral membrane protein	*Streptomyces coelicolor*	1-228	97.80
TRAAK	*Homo sapiens*	9-227	97.77
Copper-binding protein	*Methanosarcina mazei*	40-190	97.75
CG1246-PB	*Drosophila melanogaster*	2-226	97.71
Prolipoprotein diacylglyceryl transferase	*Mycobacterium tuberculosis*	47-229	97.67
C36H8.1	*Caenorhabditis elegans*	42-227	97.63
Cytochrome o ubiquinol oxidase subunit IV	*Bartonella henselae*	17-221	97.55
F47B8.5	*Caenorhabditis elegans*	113-228	97.49
Pannexin 2	*Homo sapiens*	51-229	97.45
R09E10.9	*Caenorhabditis elegans*	69-227	97.40
F47B8.5	*Caenorhabditis elegans*	103-229	97.37
CG17010-PA, isoform A	*Drosophila melanogaster*	110-223	97.36
SH3-domain binding protein 1	*Homo sapiens*	1-228	97.35
Repellent protein 1 precursor		118-228	97.33
Polygalacturonase	*Arabidopsis thaliana*	109-227	97.27
Polygalacturonase	*Arabidopsis thaliana*	119-229	97.24
Repellent protein 1 precursor		114-220	97.24
TonB family protein	*Nostoc punctiforme*	37-225	97.21
Eukaryotic translation initiation factor 3	*Homo sapiens*	120-204	97.14
Eukaryotic translation initiation factor 3	*Homo sapiens*	137-220	97.14
Cytochrome o ubiquinol oxidase subunit IV	*Bartonella henselae*	32-225	97.10
R09E10.9	*Caenorhabditis elegans*	13-227	97.10
Armadillo repeat containing, X-linked 4	*Homo sapiens*	105-228	97.05
Prolipoprotein diacylglyceryl transferase	*Mycobacterium tuberculosis*	103-229	97.04
Nischarin	*Mus musculus*	107-217	97.01

Transporter	*Arabidopsis thaliana*	111-228	96.96
Nischarin	*Mus musculus*	108-229	96.95
SphingoMyelin Synthase family member (sms-1)	*Caenorhabditis elegans*	109-203	96.90
CG1246-PB	*Drosophila melanogaster*	15-219	96.86
Related to CSR1 - phosphatidylinositol transfer protein		111-219	96.83
PPARgamma constitutive coactivator 1	*Homo sapiens*	102-226	96.79
Eukaryotic translation initiation factor 3 subunit	*Homo sapiens*	132-203	96.73
Gas vesicle protein L	*Frankia alni*	111-224	96.71
SphingoMyelin Synthase family member (sms-1)	*Caenorhabditis elegans*	109-229	96.70
Eukaryotic translation initiation factor 3 subunit	*Homo sapiens*	135-207	96.68
Eukaryotic translation initiation factor 3, subunit 9	*Mus musculus*	105-226	96.66
Transporter	*Arabidopsis thaliana*	107-229	96.65
C36H8.1	*Caenorhabditis elegans*	3-223	96.65
CG4875-PB, isoform B	*Drosophila melanogaster*	80-227	96.65
C05E11.1	*Caenorhabditis elegans*	29-227	96.64
K09E4.6	*Caenorhabditis elegans*	90-209	96.58
BLASTP/PSIBLAST			
Mitochondria Localisation Sequence		106-210	1.51e-04
Ribonuclease E; Reviewed		91-219	1.16e-12
Ehrlichia tandem repeat (Ehrlichia_rpt)		102-218	1.43e-05
Terminal organelle assembly protein TopJ		102-218	2.26e-04
Bv80/Bb-1, partial	*Babesia bovis*	98-226	3e-15
Bv80/Bb-1, partial	*Babesia bovis*	98-224	4e-13
PSIBLAST			
Protein B602L, partial	*Columba livia*	98-222	4e-15
85 kDa protein **(2)**	*Babesia bovis*	102-221	2e-14, 5e-14
Bv80, partial **(2)**	*Babesia bovis*	101-227	2e-12, 1e-10
Cell surface protein, partial **(2)**	*Bacillus thuringiensis*	89-218	3e-12-6e-12
S-layer protein precursor	*Bacillus thuringiensis*	89-218	3e-12
Response regulator receiver domain protein (CheY-like)	*Nodularia spumigena*	134-227	8e-12
Bv80/Bb-1, partial	*Babesia bovis*	101-226	7e-11

Bv80, partial **(2)**	*Babesia bovis*	102-227	7e-11, 4e-09
Cell surface protein	*Bacillus thuringiensis*	96-218	9e-11
B602L, partial	*African swine fever virus*	86-221	2e-10
IgA1 protease precursor	*Erwinia billingiae*	97-224	4e-10
BV80 merozoite protein	*Babesia bovis*	129-221	1e-09
Bv80, partial	*Babesia bovis*	98-206	2e-09
Pathogenicity protein	*Weissella ceti*	79-217	2e-09
Pullulanase, type I	*Lachnospiraceae bacterium*	72-226	3e-09
9RL protein	*African swine fever virus*	87-226	3e-09
9RL	*African swine fever virus*	98-222	3e-09
B602L, partial	*African swine fever virus*	86-219	4e-09
LdOrf-129 peptide	*Lymantria dispar multiple nucleopolyhedrovirus*	96-182	4e-09
9RL	*African swine fever virus*	98-215	4e-09
Cell division protein FtsY	*Filamentous cyanobacterium ESFC-1*	107-215	4e-09
B602L protein	*African swine fever virus*	88-215	5e-09
Snaclec 3	*Toxocara canis*	97-221	5e-09
9RL	*African swine fever virus*	97-226	2e-08
FHA domain containing protein	*Arthrospira platensis*	89-218	2e-08
Liver stage antigen 3	*Plasmodium falciparum*	102-222	3e-08
85 kDa protein	*Babesia bovis*	107-185	5e-08
B602L protein	*African swine fever virus*	88-222	6e-08
B602L	*African swine fever virus*	108-227	6e-08
ORF-132 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	122-214	6e-08
Glutamate/valine-rich protein	*Natronorubrum sulfidifaciens*	104-222	9e-08
Type I restriction modification protein	*Mycoplasma pneumoniae*	99-174	9e-08
B602L protein	*African swine fever virus*	108-219	1e-07
Bv80/Bb-1, partial	*Babesia bovis*	103-190	2e-07
B602L protein	*African swine fever virus*	87-219	2e-07
B602L protein	*African swine fever virus*	103-215	2e-07
B602L **(2)**	*African swine fever virus*	86-215	2e-07, 1e-05
pB602L **(2)**	*African swine fever virus*	88-221	3e-07, 4e-

			07
pB602L	*African swine fever virus*	99-215	4e-07
B602L protein **(2)**	*African swine fever virus*	87-203	5e-07, 6e-07
Cell division protein FtsK **(18)**	*Burkholderia pseudomallei*	99-221	6e-07-0.001
9RL protein **(2)**	*African swine fever virus*	87-199	6e-07, 9e-06
ftsK/SpoIIIE family protein **(2)**	*Burkholderia pseudomallei*	99-221	7e-07, 1e-04
B602L protein **(6)**	*African swine fever virus*	93-199	1e-06-1e-05
B602L protein	*African swine fever virus*	87-207	1e-06
9RL	*African swine fever virus*	96-207	1e-06
pB602L	*African swine fever virus*	88-223	2e-06
9RL	*African swine fever virus*	96-199	2e-06
pB602L, partial	*African swine fever virus*	87-205	2e-06
Cellulosomal scaffoldin anchoring protein	*Trichomonas vaginalis*	98-219	5e-06
kxYKxGKxW signal peptide domain protein	*Streptococcus mitis*	104-221	5e-06
Cell division protein FtsK **(6)**	*Burkholderia pseudomallei*	105-221	7e-06-4e-04
Trans-sialidase	*Trypanosoma cruzi*	97-219	1e-05
B602L protein	*African swine fever virus*	96-199	2e-05
GH24581	*Drosophila grimshawi*	125-219	2e-05
Cell division protein FtsK	*Burkholderia sp. TSV202*	106-221	2e-05
Cell division protein FtsK	*Burkholderia sp. MSHR44*	105-221	2e-05
9RL protein, partial	*African swine fever virus*	103-223	3e-05
Proteoglycan 4	*Pteropus alecto*	103-218	4e-05
Peptidase	*Actinomyces sp. oral*	100-222	6e-05
Cell division protein FtsK **(3)**	*Burkholderia pseudomallei*	106-221	1e-04-0.003
Peptidase	*Actinomyces viscosus*	105-222	2e-04
DNA translocase FtsK domain protein	*Burkholderia pseudomallei*	102-221	2e-04
DNA translocase FtsK	*Burkholderia pseudomallei*	99-221	3e-04
9RL protein, partial	*African swine fever virus*	103-219	3e-04
B602L protein	*African swine fever virus*	93-223	4e-04

Multispecies: cell division protein FtsK	*Burkholderia*	99-221	5e-04
Cell divisionftsk/spoiiie	*Burkholderia pseudomallei*	105-221	5e-04
Ribonuclease E	*Marinomonas sp. S3726*	95-221	0.001
DNA translocase FtsK	*Ralstonia solanacearum*	95-220	0.002
Multispecies: nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase	*Streptomyces*	102-218	0.004
Trans-sialidase	*Trypanosoma cruzi*	97-218	0.004
I-TASSER			
Survival motor neuron protein **(3)**	*Homo sapiens*		1.17, 1.33, 1.85
Accumulation associated protein	*Staphylococcus epidermidis*		1.42
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		1.99
Myc box dependent interacting protein 1 **(2)**	*Homo sapiens*		1.03, 2.03
Chitinase 60	*Moritella marina*		1.17
Major capsid protein	*Synechococcus phage Syn5*		1.75
Anosmin 1	*Homo sapiens*		1.08
Survival motor neuron protein	*Homo sapiens*		0.867
Predict Protein			
Cytoplasm			
Proteoglycan 4 **(11)**	*Mus musculus*		2e-11-0.003
Titin **(20)**	*Mus musculus*		9e-10-0.050
Atome2			
Na(+)/H(+) exchange regulatory cofactor NHE-RF1	*Homo sapiens*		76.96
26S proteasome non-ATPase regulatory subunit 4 ((poly)ubiquitin binding region)	*Homo sapiens*		57.25
ADP-ribosylation factor binding protein GGA1	*Homo sapiens*		45.01
Ribosome-interacting GTPase 1	*Saccharomyces cerevisiae*		25.85
EspA	*Escherichia coli*		22.74

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of

10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXIV. *Utterbackia imbecillis* H-*ORF* sequence 3 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		71-74	99.33
TIGR04294 prepilin-type processing-associated H-X9-DG domain		25-30	99.27
TIGR01167 LPXTG cell wall anchor domain		49-64	98.98
G protein-coupled receptor 152	*Homo sapiens*	2-192	99.04
Procyclic acidic repetitive protein (PARP)		62-163	98.98
G protein-coupled receptor 152	*Homo sapiens*	2-187	98.79
Procyclic acidic repetitive protein (PARP)		66-167	98.78
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	92-193	98.76
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	96-192	98.71
T06A4.1b	*Caenorhabditis elegans*	84-192	98.55
Related to CSR1 - phosphatidylinositol transfer protein		94-191	98.53
Related to CSR1 - phosphatidylinositol transfer protein		94-192	98.53
T06A4.1b	*Caenorhabditis elegans*	98-179	98.47
Y22D7AR.1	*Caenorhabditis elegans*	16-192	98.44
CG1468-PA	*Drosophila melanogaster*	83-192	98.43
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	92-192	98.40
CG1468-PA	*Drosophila melanogaster*	98-195	98.39
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		43-45	97.70
T10G3.1	*Caenorhabditis elegans*	72-193	98.30
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	96-193	98.27
R06C7.4	*Caenorhabditis elegans*	93-191	98.26
T10G3.1	*Caenorhabditis elegans*	92-189	98.25
R06C7.4	*Caenorhabditis elegans*	73-183	98.19

CG9029-PA	*Drosophila melanogaster*	78-186	98.18
V-set and immunoglobulin domain containing 1	*Mus musculus*	51-187	98.15
Y22D7AR.1	*Caenorhabditis elegans*	71-193	98.15
CG9029-PA	*Drosophila melanogaster*	72-183	98.08
T04F8.8	*Caenorhabditis elegans*	105-170	98.06
V-set and immunoglobulin domain containing 1	*Mus musculus*	93-183	98.04
T04F8.8	*Caenorhabditis elegans*	97-156	98.01
F56B6.4a	*Caenorhabditis elegans*	96-193	98.01
F49B2.6	*Caenorhabditis elegans*	96-190	98.01
Nucleolar protein 3	*Homo sapiens*	77-187	97.99
F47B8.5	*Caenorhabditis elegans*	26-189	97.95
T14A8.2	*Caenorhabditis elegans*	56-192	97.90
F56B6.4a	*Caenorhabditis elegans*	96-192	97.90
F49B2.6	*Caenorhabditis elegans*	94-192	97.89
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	94-184	97.87
TIGR03501 GlyGly-CTERM domain		71-79	97.05
Protein with Tau-Like repeats family member (ptl-1)	*Caenorhabditis elegans*	92-190	97.77
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	44-190	97.73
Solute carrier family 16, member 2	*Homo sapiens*	96-182	97.71
Repellent protein 1 precursor		96-185	97.70
Repellent protein 1 precursor		96-189	97.70
C09F9.2	*Caenorhabditis elegans*	94-187	97.67
Solute carrier family 16, member 2	*Homo sapiens*	96-192	97.58
Diacylglycerol kinase kappa	*Homo sapiens*	95-193	97.56
Nucleolar protein 3	*Homo sapiens*	94-180	97.56
C09F9.2	*Caenorhabditis elegans*	93-192	97.53
V-set and immunoglobulin domain containing 1	*Homo sapiens*	59-192	97.48
Diacylglycerol kinase kappa	*Homo sapiens*	97-193	97.43
Nucleolar protein 3	*Mus musculus*	98-188	97.35
Sperm equatorial segment protein 1	*Mus musculus*	98-192	97.33
Cytochrome o ubiquinol oxidase subunit IV	*Bartonella henselae*	25-190	97.32
C15C8.1	*Caenorhabditis elegans*	47-192	97.29
V-set and immunoglobulin domain containing 1	*Homo sapiens*	94-184	97.29
Integral membrane protein	*Streptomyces coelicolor*	33-192	97.26
Sperm equatorial segment protein 1	*Mus musculus*	89-189	97.22
Y51F10.4a	*Caenorhabditis elegans*	12-192	97.22

cyclin K	*Mus musculus*	7-185	97.22
K09E4.6	*Caenorhabditis elegans*	54-174	97.16
F47B8.5	*Caenorhabditis elegans*	98-191	97.16
Protease inhibitor Kazal-type	*Nitrosopumilus maritimus*	97-189	97.08
K09E4.6	*Caenorhabditis elegans*	110-186	97.08
Cytochrome o ubiquinol oxidase subunit IV	*Bartonella henselae*	10-192	97.06
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	94-187	96.99
TRAAK	*Homo sapiens*	2-189	96.97
Prolipoprotein diacylglyceryl transferase	*Mycobacterium tuberculosis*	43-192	96.97
Procyclic acidic repetitive protein (PARP)		99-161	96.94
BLASTP/PSIBLAST			
Mitochondria Localisation Sequence		91-178	1.83e-04
Ribonuclease E; Reviewed		94-190	1.68e-08
Ehrlichia tandem repeat (Ehrlichia_rpt)		95-186	2.45e-04
Terminal organelle assembly protein TopJ		102-192	1.00e-03
Bv80/Bb-1, partial	*Babesia bovis*	98-194	5e-12
Protein B602L, partial	*Columba livia*	98-191	3e-11
Bv80/Bb-1, partial	*Babesia bovis*	98-190	3e-11
PSIBLAST			
S-layer protein precursor	*Bacillus thuringiensis*	89-186	1e-12
Cell surface protein **(3)**	*Bacillus thuringiensis*	89-186	1e-12-3e-12
Response regulator receiver domain protein (CheY-like)	*Nodularia spumigena*	98-195	4e-12
85 kDa merozoite protein	*Babesia bovis*	102-190	1e-11
85 kDa protein **(2)**	*Babesia bovis*	98-190	2e-11, 4e-10
BV80 merozoite protein	*Babesia bovis*	102-189	7e-10
LdOrf-129 peptide	*Lymantria dispar multiple nucleopolyhedrovirus*	96-184	8e-10
85 kDa protein	*Babesia bovis*	99-174	8e-10
Outer membrane autotransporter barrel domain-containing protein	*Escherichia coli*	107-182	3e-09
ORF-132 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	96-184	5e-09
Bv80/Bb-1, partial	*Babesia bovis*	98-194	5e-09

Bv80/Bb-1, partial **(3)**	*Babesia bovis*	98-195	1e-08-4e-07
Bv80/Bb-1, partial	*Babesia bovis*	108-194	1e-08
IgA1 protease precursor	*Erwinia billingiae*	97-189	6e-08
Type I restriction modification protein	*Mycoplasma pneumoniae*	99-174	6e-08
Bv80, partial	*Babesia bovis*	98-190	2e-07
Ribonuclease E	*Marinomonas sp. S3726*	95-191	2e-06
B602L, partial **(4)**	*African swine fever virus*	82-189	2e-06-2e-04
Restriction endonuclease, S subunit	*Mycoplasma pneumoniae*	98-174	2e-06
B602L, partial	*African swine fever virus*	82-190	2e-06
Pullulanase, type I	*Lachnospiraceae bacterium*	97-183	3e-06
9RL protein	*African swine fever virus*	87-187	4e-06
B602L protein **(2)**	*African swine fever virus*	88-188	6e-06, 2e-05
Type I restriction modification protein	*Mycoplasma pneumoniae*	98-170	7e-06
Cell division protein FtsY	*Filamentous cyanobacterium*	97-183	8e-06
Restriction endonuclease, S subunit	*Mycoplasma pneumoniae*	98-170	8e-06
B602L protein, partial **(6)**	*African swine fever virus*	87-189	1e-05-0.001
9RL	*African swine fever virus*	97-188	2e-05
B602L	*African swine fever virus*	87-188	2e-05
Snaclec 3	*Toxocara canis*	97-189	2e-05
B602L protein	*African swine fever virus*	87-187	3e-05
B602L protein	*African swine fever virus*	87-191	3e-05
9RL protein **(2)**	*African swine fever virus*	87-189	3e-05, 9e-05
B602L, partial	*African swine fever virus*	82-178	3e-05
B602L **(3)**	*African swine fever virus*	86-189	5e-05-0.002
9RL	*African swine fever virus*	97-190	6e-05
Central variable region protein	*African swine fever virus*	82-183	6e-05
Central variable region protein	*African swine fever virus*	82-194	8e-05
B602L protein **(6)**	*African swine fever virus*	93-189	9e-05-7e-04
Cell divisionftsk/spoiiie **(3)**	*Burkholderia pseudomallei*	95-189	9e-05-

			0.003
Cell division protein FtsK	*Burkholderia sp. TSV202*	99-189	1e-04
Cell division protein FtsK **(2)**	*Burkholderia pseudomallei*	99-189	1e-04, 2e-04
B602L protein	*African swine fever virus*	87-189	2e-04
9RL	*African swine fever virus*	96-190	2e-04
Type I restriction modification protein, partial	*Mycoplasma pneumoniae*	130-191	2e-04
orf-126 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	94-164	2e-04
Cell division protein FtsK	*Burkholderia pseudomallei*	99-192	2e-04
Cell division FtsK/SpoIIIE	*Burkholderia pseudomallei*	99-189	2e-04
GH24581	*Drosophila grimshawi*	99-187	2e-04
9RL protein	*African swine fever virus*	87-183	2e-04
pB602L, partial	*African swine fever virus*	87-194	3e-04
9RL protein	*African swine fever virus*	87-194	3e-04
Cell division protein FtsK	*Burkholderia sp. MSHR44*	105-189	4e-04
pB602L	*African swine fever virus*	99-188	4e-04, 5e-04
Peptidase	*Actinomyces sp. oral*	105-188	4e-04
B602L protein	*African swine fever virus*	103-188	5e-04
B602L, partial	*African swine fever virus*	86-194	6e-04
B602L protein	*African swine fever virus*	96-190	7e-04
kxYKxGKxW signal peptide domain protein	*Streptococcus mitis*	104-189	8e-04
pB602L	*African swine fever virus*	88-194	0.001
pB602L	*African swine fever virus*	88-189	0.001
9RL	*African swine fever virus*	97-193	0.001
B602L protein	*African swine fever virus*	94-189	0.001
B602L protein	*African swine fever virus*	87-183	0.001
Proteoglycan 4	*Tupaia chinensis*	105-149	0.002
CG3108	*Drosophila melanogaster*	102-191	0.002
B602L protein	*African swine fever virus*	103-194	0.002
9RL protein, partial	*African swine fever virus*	103-191	0.002
Peptidase	*Actinomyces viscosus*	96-190	0.002
B602L protein **(2)**	*African swine fever virus*	98-193	0.003
B602L, partial	*African swine fever virus*	86-187	0.003
9RL	*African swine fever virus*	97-190	0.004

Cellulosomal scaffoldin anchoring protein	*Trichomonas vaginalis*	98-187	0.005
B602L protein	*African swine fever virus*	87-187	0.005
Cell division protein FtsK	*Burkholderia pseudomallei*	95-189	0.005
I-TASSER			
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		2.10
Myc box dependent interacting protein 1	*Homo sapiens*		1.55
60S ribosomal protein L1	*Saccharomyces cerevisiae*		1.49
Survival motor neuron protein	*Homo sapiens*		1.37
Double-stranded RNA-specific editase 1	*Rattus norvegicus*		1.34
Major capsid protien	*Synechococcus phage Syn5*		1.28
gp7	*Salmonella phage epsilon15*		1.32
Polymeric-immunoglobulin receptor	*Homo sapiens*		1.24
Long tail fiber protein P37	*Enterobacteria phage*		1.30
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		0.922
phospholipase C beta	*Meleagris gallopavo*		0.704
Interferon-induced guanylate-binding protein 1	*Homo sapiens*		0.664
Dynamin family protein	*Nostoc punctiforme*		0.632
RhUL123	*Macacine herpesvirus 3*		0.617
Tyrosine-protein kinase Fes/Fps	*Homo sapiens*		0.605
Formin-binding protein 1	*Homo sapiens*		0.595
Cdc42-interacting protein 4	*Homo sapiens*		0.586
Apolipoprotein A-IV	*Homo sapiens*		0.582
Predict Protein			
Protein binding		69-71, 74	
Cytoplasm			
Diacylglycerol kinase kappa **(3)**	*Homo sapiens*		4e-08-4e-07
Proteoglycan 4 **(2)**	*Homo sapiens*		0.25, 0.58
Proteoglycan 4 **(11)**	*Mus musculus*		7e-08-0.096
Atome2			
Transcriptional activator rfaH	*Escherichia coli*		76.29
80 kDa MCM3-associated protein	*Homo sapiens*		71.48
Ubiquitin carboxyl-terminal hydrolase 28	*Homo sapiens*		70.13

Ribosome-interacting GTPase 1	*Saccharomyces cerevisiae*		64.19
Delta sleep inducing peptide immunoreactive peptide	*Sus scrofa*		58.92
26S proteasome non-ATPase regulatory subunit 4	*Homo sapiens*		56.57
Nucleoprotein	*Andes virus*		39.59
Clathrin heavy chain 1	*Bos taurus*		7.88

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXV. *Utterbackia imbecillis* H-*ORF* sequence 4 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		71-74	99.28
TIGR04294 prepilin-type processing-associated H-X9-DG domain		25-30	99.14
TIGR01167 LPXTG cell wall anchor domain		49-64	98.89
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		43-45	97.50
Copper-binding protein	*Methanosarcina mazei*	33-174	97.89
TIGR03501 GlyGly-CTERM domain		71-79	96.95
G protein-coupled receptor 152	*Homo sapiens*	2-175	96.79
CG12522-PA	*Drosophila melanogaster*	89-155	96.42
CG12522-PA	*Drosophila melanogaster*	97-162	96.25
Copper-binding protein	*Methanosarcina mazei*	96-170	96.18
K09E4.6	*Caenorhabditis elegans*	96-158	95.95
Eukaryotic translation initiation factor 3, subunit 9	*Mus musculus*	95-162	95.89
T14A8.2	*Caenorhabditis elegans*	56-166	95.87
K09E4.6	*Caenorhabditis elegans*	54-163	95.54
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	96-159	95.46

Eukaryotic translation initiation factor 3, subunit 9	*Mus musculus*	97-163	95.42
TIGR00756 pentatricopeptide repeat domain		36-43	94.09
Armadillo repeat containing, X-linked 4	*Homo sapiens*	91-168	95.07
Prolipoprotein diacylglyceryl transferase	*Frankia alni*	96-166	94.63
Cytochrome o ubiquinol oxidase subunit IV	*Bartonella henselae*	10-168	94.30
DumPY: shorter than wild-type family member (dpy-10)	*Caenorhabditis elegans*	23-88	94.09
Armadillo repeat containing, X-linked 4	*Homo sapiens*	102-172	93.67
Prolipoprotein diacylglyceryl transferase	*Mycobacterium tuberculosis*	44-164	93.57
Nischarin	*Homo sapiens*	98-161	93.34
MSP1_C: Merozoite surface protein 1 (MSP1) C-terminus		92-167	93.31
G protein-coupled receptor 152	*Homo sapiens*	2-166	93.23
30S ribosomal protein S3P	*Methanosarcina mazei*	97-166	92.92
TonB family protein	*Nostoc punctiforme*	30-166	92.81
BLIstered cuticle family member (bli-2)	*Caenorhabditis elegans*	34-89	92.48
30S ribosomal protein S3P	*Methanosarcina mazei*	91-166	92.19
Y51F10.4a	*Caenorhabditis elegans*	12-168	91.98
Membralin isoform 1	*Homo sapiens*	9-168	91.79
SphingoMyelin Synthase family member (sms-1)	*Caenorhabditis elegans*	97-167	91.58
Nischarin	*Mus musculus*	98-167	91.53
F57B1.3	*Caenorhabditis elegans*	29-82	91.07
SphingoMyelin Synthase family member (sms-1)	*Caenorhabditis elegans*	94-162	90.84
PPARgamma constitutive coactivator 1	*Homo sapiens*	93-166	90.47
T10E10.5	*Caenorhabditis elegans*	34-89	90.32
CG32372-PA	*Drosophila melanogaster*	95-161	90.19
Nischarin	*Homo sapiens*	98-162	90.16
PPARgamma constitutive coactivator 1	*Homo sapiens*	96-162	90.07
F33A8.9	*Caenorhabditis elegans*	32-83	90.06
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	34-85	90.02
CG4875-PB, isoform B	*Drosophila melanogaster*	78-168	89.93
CG32372-PA	*Drosophila melanogaster*	95-168	89.52
DumPY: shorter than wild-type family member (dpy-14)	*Caenorhabditis elegans*	42-89	89.50
SH3 type 3 domain-containing protein	*Nostoc punctiforme*	31-160	89.35
Cell wall structural complex MreBCD transmembrane component MreC	*Escherichia coli*	91-166	89.30
Inner membrane protein translocase component YidC	*Streptomyces coelicolor*	10-166	89.12

COLlagen family member (col-102)	*Caenorhabditis elegans*	47-83	88.90
Binding	*Arabidopsis thaliana*	102-176	88.61
CG12316-PB, isoform B	*Drosophila melanogaster*	98-162	88.48
CG12316-PA, isoform A	*Drosophila melanogaster*	98-162	88.48
Cell division protein FtsY	*Frankia alni*	75-161	88.40
CG12316-PA, isoform A	*Drosophila melanogaster*	94-167	88.28
CG12316-PB, isoform B	*Drosophila melanogaster*	94-167	88.28
Binding	*Arabidopsis thaliana*	106-160	88.14
Prolipoprotein diacylglyceryl transferase	*Mycobacterium tuberculosis*	96-166	88.10
Y54E10BL.2	*Caenorhabditis elegans*	44-89	87.75
DumPY: shorter than wild-type family member (dpy-3)	*Caenorhabditis elegans*	34-85	87.65
SH3-domain binding protein 1	*Homo sapiens*	3-161	87.46
Nischarin	*Mus musculus*	92-168	87.44
Eukaryotic translation initiation factor 3, subunit 5 epsilon	*Homo sapiens*	96-180	87.42
COLlagen family member (col-101)	*Caenorhabditis elegans*	38-89	87.38
BLASTP/PSIBLAST			
bifunctional 2',3'-cyclic nucleotide 2'-phosphodiesterase/3'-nucleotidase precursor protein; Reviewed		88-156	3.57e-04
Bv80/Bb-1, partial **(2)**	*Babesia bovis*	99-158	1e-11, 1e-10
Bv80/Bb-1, partial	*Babesia bovis*	98-172	2e-11
Bv80/Bb-1, partial	*Babesia bovis*	89-158	4e-11
85 kDa protein	*Babesia bovis*	99-158	1e-10
Bv80, partial **(3)**	*Babesia bovis*	98-158	1e-10-4e-10
Cell surface protein, partial	*Bacillus thuringiensis*	95-175	4e-10
S-layer protein precursor	*Bacillus thuringiensis*	95-175	4e-10
Protein B602L, partial	*Columba livia*	98-173	8e-10
85 kDa protein	*Babesia bovis*	98-158	2e-09
Cell surface protein	*Bacillus thuringiensis*	96-175	4e-09
85 kDa merozoite protein	*Babesia bovis*	102-158	5e-09
ORF-132 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	96-175	1e-08
LdOrf-129 peptide	*Lymantria dispar multiple*	96-156	3e-08

	nucleopolyhedrovirus		
GH24581	*Drosophila grimshawi*	88-176	2e-06
Type I restriction modification protein	*Mycoplasma pneumoniae*	98-174	2e-06
Restriction endonuclease, S subunit	*Mycoplasma pneumoniae*	98-170	9e-06
Type I restriction modification protein	*Mycoplasma pneumoniae*	98-170	1e-05
Restriction endonuclease, S subunit	*Mycoplasma pneumoniae*	98-173	1e-05
PSIBLAST			
Cell surface protein	*Bacillus thuringiensis*	96-175	5e-09
Bv80, partial **(3)**	*Babesia bovis*	98-158	4e-08-1e-05
Outer membrane autotransporter barrel domain-containing protein	*Escherichia coli*	95-158	3e-07
Response regulator receiver domain protein (CheY-like)	*Nodularia spumigena*	92-158	3e-07
orf-126 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	96-158	6e-07
Central variable region protein	*African swine fever virus*	82-160	2e-06
Bv80, partial **(3)**	*Babesia bovis*	102-157	3e-06-2e-05
Bv80, partial	*Babesia bovis*	102-158	5e-06
9RL protein	*African swine fever virus*	87-162	1e-05
Central variable region protein	*African swine fever virus*	82-158	1e-05
B602L protein **(8)**	*African swine fever virus*	95-160	2e-05-4e-04
9RL, partial	*African swine fever virus*	82-160	3e-05
B602L **(4)**	*African swine fever virus*	82-158	4e-05-0.003
Bv80/Bb-1 **(5)**	*Babesia bovis*	106-157	5e-05-1e-04
B602L, partial	*African swine fever virus*	82-151	7e-05
B602L protein	*African swine fever virus*	88-160	1e-04
B602L protein	*African swine fever virus*	87-157	1e-04
9RL protein **(2)**	*African swine fever virus*	87-174	1e-04, 0.004
9RL, partial	*African swine fever virus*	82-158	2e-04
9RL protein, partial	*African swine fever virus*	87-160	2e-04

Peptidase	*Actinoplanes sp. SE50/110*	109-157	2e-04
B602L protein	*African swine fever virus*	87-159	3e-04
9RL **(2)**	*African swine fever virus*	97-162	3e-04, 0.001
CG3108	*Drosophila melanogaster*	95-168	3e-04
Type I restriction modification protein, partial	*Mycoplasma pneumoniae*	100-158	3e-04
9RL protein	*African swine fever virus*	87-157	4e-04
B602L protein	*African swine fever virus*	88-158	4e-04
Central variable region protein	*African swine fever virus*	82-174	5e-04, 7e-04
9RL protein **(2)**	*African swine fever virus*	82-174	6e-04, 0.001
B602L protein **(3)**	*African swine fever virus*	87-158	7e-04, 0.002
B602L protein	*African swine fever virus*	88-174	8e-04
B602L protein **(4)**	*African swine fever virus*	88-157	9e-04-0.004
B602L, partial	*African swine fever virus*	82-155	0.001
pB602L **(3)**	*African swine fever virus*	82-157	0.001
9RL protein, partial	*African swine fever virus*	97-174	0.001
B602L protein	*African swine fever virus*	88-162	0.001
pB602L, partial	*African swine fever virus*	87-171	0.001
Cell division protein FtsK	*Ralstonia solanacearum*	89-181	0.001
B602L protein	*African swine fever virus*	98-162	0.002
pB602L	*African swine fever virus*	82-157	0.002
GG21615	*Drosophila erecta*	87-160	0.002
B602L protein	*African swine fever virus*	88-166	0.002
GI15252	*Drosophila mojavensis*	91-177	0.002
Involucrin repeat protein	*Ophiostoma piceae*	96-157	0.002
B602L protein	*African swine fever virus*	87-155	0.003
9RL	*African swine fever virus*	97-162	0.003
Type IV secretion protein Rhs, partial	*Nocardioides sp. URHA0020*	98-172	0.004
9RL protein, partial	*African swine fever virus*	96-174	0.004
EFG1p-dependent transcript 1 protein	*Candida albicans*	113-174	0.004
I-TASSER			
Type I hyperactive antifreeze protein	*Pseudopleuronectes*		1.07, 2.30

	americanus		
Accumulation associated protein	*Staphylococcus epidermidis*		1.44
Cellulosomal scaffoldin adaptor protein B	*Acetivibrio cellulolyticus*		1.00
Internalin K	*Listeria monocytogenes*		1.04
Myc box dependent interacting protein 1	*Homo sapiens*		1.41
30S ribosomal protein S10	*Thermus thermophilus*		1.09
Survival motor neuron protein	*Homo sapiens*		1.24, 1.26
Tropomyosin	*Oryctolagus cuniculus*		1.48
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		0.821
Dynamin family protein	*Nostoc punctiforme*		0.644
Phospholipase C beta	*Meleagris gallopavo*		0.629
RhUL123	*Macacine herpesvirus 3*		0.608
Interferon-induced guanylate-binding protein 1	*Homo sapiens*		0.600
Tyrosine-protein kinase Fes/Fps	*Homo sapiens*		0.599
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-3	*Homo sapiens*		0.594
LEOA	*Escherichia coli*		0.580
Apolipoprotein A-IV	*Homo sapiens*		0.575
Predict Protein			
Protein binding		30, 72, 74, 92, 179, 183	
Polynucleotide binding		27	
Cytoplasm			
Proteoglycan 4 **(11)**	*Mus musculus*		7e-07-0.18
Protein psiA **(6)**	*Dictyostelium discoideum*		0.047-0.21
Atome2			
Ubiquitin carboxyl-terminal hydrolase 28	*Homo sapiens*		82.82
Hydrophilic protein; has cysteine rich putative zinc finger essential for function; Vps27p	*Saccharomyces cerevisiae*		62.08
80 kDa MCM3-associated protein	*Homo sapiens*		58.91
Delta sleep inducing peptide immunoreactive peptide	*Sus scrofa*		56.98
12 kDa heat shock protein	*Saccharomyces cerevisiae*		55.06
Nucleoprotein	*Andes virus*		52.84
Antifreeze protein isoform HPLC6	*Pseudopleuronectes*		47.80

	americanus		
Sensory rhodopsin II	*Natronomonas pharaonis*		46.16
Zinc resistance-associated protein	*Salmonella enterica*		38.16
Autocrine motility factor receptor, isoform 2	*Homo sapiens*		28.24

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXVI. *Utterbackia imbecillis* H-*ORF* sequences 5 & 6 function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		71-74	99.19
TIGR04294 prepilin-type processing-associated H-X9-DG domain		13-15	99.14
TIGR01167 LPXTG cell wall anchor domain		97-99	98.86
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		43-45	97.45
TIGR03501 GlyGly-CTERM domain		55-70	96.72
TIGR00756 pentatricopeptide repeat domain		48-57	94.26
CG12522-PA	*Drosophila melanogaster*	96-143	92.71
DumPY: shorter than wild-type family member (dpy-10)	*Caenorhabditis elegans*	23-88	92.56
CG12522-PA	*Drosophila melanogaster*	94-147	91.67
G protein-coupled receptor 152	*Homo sapiens*	2-146	88.69
F57B1.3	*Caenorhabditis elegans*	29-89	88.16
COLlagen family member (col-102)	*Caenorhabditis elegans*	47-89	87.04
Y54E10BL.2	*Caenorhabditis elegans*	44-89	86.74
DumPY: shorter than wild-type family member (dpy-14)	*Caenorhabditis elegans*	42-89	86.56
T10E10.5	*Caenorhabditis elegans*	34-89	86.53
COLlagen family member (col-69)	*Caenorhabditis elegans*	32-89	86.49
C12D8.8	*Caenorhabditis elegans*	41-89	86.44

BLIstered cuticle family member (bli-2)	*Caenorhabditis elegans*	38-92	85.59
DumPY: shorter than wild-type family member (dpy-9)	*Caenorhabditis elegans*	43-89	83.99
ROLler: helically twisted, animals roll when moving family member (rol-8)	*Caenorhabditis elegans*	44-89	83.51
W08D2.6	*Caenorhabditis elegans*	40-91	83.28
F54B11.1	*Caenorhabditis elegans*	31-89	82.65
COLlagen family member (col-2)	*Caenorhabditis elegans*	43-89	81.76
COLlagen family member (col-145)	*Caenorhabditis elegans*	46-89	81.14
F15A2.1	*Caenorhabditis elegans*	41-89	80.78
COLlagen family member (col-101)	*Caenorhabditis elegans*	38-89	80.34
DumPY: shorter than wild-type family member (dpy-3)	*Caenorhabditis elegans*	34-89	79.98
COLlagen family member (col-166)	*Caenorhabditis elegans*	33-89	79.95
T10E10.2	*Caenorhabditis elegans*	34-89	79.75
COLlagen family member (col-162)	*Caenorhabditis elegans*	47-89	79.50
F11G11.11	*Caenorhabditis elegans*	38-89	79.38
COLlagen family member (col-176)	*Caenorhabditis elegans*	41-89	79.11
COLlagen family member (col-14)	*Caenorhabditis elegans*	15-89	79.11
C30F2.1	*Caenorhabditis elegans*	45-89	78.99
C44C10.1	*Caenorhabditis elegans*	34-89	78.67
DumPY: shorter than wild-type family member (dpy-2)	*Caenorhabditis elegans*	26-82	78.54
COLlagen family member (col-65)	*Caenorhabditis elegans*	37-82	78.49
T06E4.4	*Caenorhabditis elegans*	46-83	78.21
COLlagen family member (col-173)	*Caenorhabditis elegans*	35-89	77.98
T10E10.7	*Caenorhabditis elegans*	23-82	77.84
COLlagen family member (col-183)	*Caenorhabditis elegans*	33-89	77.60
COLlagen family member (col-75)	*Caenorhabditis elegans*	43-82	77.59
COLlagen family member (col-110)	*Caenorhabditis elegans*	43-89	77.42
T05A1.2	*Caenorhabditis elegans*	43-89	77.31
COLlagen family member (col-185)	*Caenorhabditis elegans*	43-100	76.07
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	33-89	75.48
SYNtaxin family member (syn-2)	*Caenorhabditis elegans*	42-68	75.43
F33A8.9	*Caenorhabditis elegans*	32-89	75.34
COLlagen family member (col-165)	*Caenorhabditis elegans*	46-89	75.28
F08G5.4	*Caenorhabditis elegans*	46-89	75.09
Y38C1BA.3	*Caenorhabditis elegans*	44-85	74.64
F54D1.3	*Caenorhabditis elegans*	34-82	74.50

C34F6.3	*Caenorhabditis elegans*	30-82	74.06
COLlagen family member (col-166)	*Caenorhabditis elegans*	11-89	74.01
COLlagen family member (col-51)	*Caenorhabditis elegans*	20-89	73.99
T06E4.6	*Caenorhabditis elegans*	46-93	73.78
F54B11.2	*Caenorhabditis elegans*	43-85	73.73
T10E10.1	*Caenorhabditis elegans*	34-89	73.70
F56D5.1	*Caenorhabditis elegans*	33-89	73.68
SQuaT family member (sqt-1)	*Caenorhabditis elegans*	42-89	73.49
T11F9.9	*Caenorhabditis elegans*	42-89	73.06
ZK1010.7	*Caenorhabditis elegans*	46-89	72.96
COLlagen family member (col-114)	*Caenorhabditis elegans*	33-89	72.74
C29F4.1	*Caenorhabditis elegans*	44-89	72.59
F32G8.5	*Caenorhabditis elegans*	45-90	72.30
F15H10.1	*Caenorhabditis elegans*	35-89	71.80
COLlagen family member (col-91)	*Caenorhabditis elegans*	47-89	71.67
LysM, putative peptidoglycan-binding, domain containing 3	*Mus musculus*	46-70	71.61
ROLler: helically twisted, animals roll when moving family member (rol-6)	*Caenorhabditis elegans*	47-89	71.51
Glutamate receptor, ionotropic, N-methyl D-asparate-associated protein 1	*Homo sapiens*	28-84	71.25
Glutamate receptor, ionotropic, N-methyl D-asparate-associated protein 1	*Homo sapiens*	28-84	71.25
COLlagen family member (col-120)	*Caenorhabditis elegans*	27-89	71.07
F55C10.2	*Caenorhabditis elegans*	42-89	71.05
D2023.7	*Caenorhabditis elegans*	47-89	70.90
COLlagen family member (col-137)	*Caenorhabditis elegans*	42-89	70.76
COLlagen family member (col-33)	*Caenorhabditis elegans*	35-92	70.35
COLlagen family member (col-84)	*Caenorhabditis elegans*	50-92	70.32
C09G5.5	*Caenorhabditis elegans*	47-94	70.32
Y41C4A.19	*Caenorhabditis elegans*	30-89	70.21
F38A3.1	*Caenorhabditis elegans*	42-91	70.06
DumPY: shorter than wild-type family member (dpy-8)	*Caenorhabditis elegans*	38-82	69.77
COLlagen family member (col-92)	*Caenorhabditis elegans*	46-89	69.51
Reticulon 1 isoform A	*Homo sapiens*	10-71	69.25
Y41C4A.16	*Caenorhabditis elegans*	30-92	69.20

C34F6.2	*Caenorhabditis elegans*	32-94	69.05
BLASTP/PSIBLAST			
ORF-132 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	96-159	2e-11/7e-08
LdOrf-129 peptide	*Lymantria dispar multiple nucleopolyhedrovirus*	96-149	4e-10/2e-06
85 kDa protein	*Babesia bovis*	91-146	9e-09/4e-05
Bv80/Bb-1, partial **(5)**	*Babesia bovis*	98-146	7e-08-1e-04
Bv80/Bb-1, partial	*Babesia bovis*	89-146	4e-07, 0.002
PSIBLAST			
S-layer protein precursor	*Bacillus thuringiensis*	95-156	2e-05
Cell surface protein, partial	*Bacillus thuringiensis*	95-156	2e-05
Cell surface protein	*Bacillus thuringiensis*	96-154	7e-05
Cell surface protein	*Bacillus thuringiensis*	97-154	9e-05
orf-126 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	96-142	1e-04
85 kDa protein **(2)**	*Babesia bovis*	98-146	1e-04
Bv80, partial	*Babesia bovis*	88-146	4e-04
Outer membrane autotransporter barrel domain-containing protein	*Escherichia coli*	95-154	5e-04
GH24581	*Drosophila grimshawi*	98-146	6e-04
Autotransporter protein, partial	*Escherichia coli*	90-146	7e-04
Outer membrane autotransporter barrel domain-containing protein	*Escherichia coli*	90-146	8e-04
GG21615	*Drosophila erecta*	87-156	0.001
Bv80, partial	*Babesia bovis*	98-146	0.002
Proteoglycan 4	*Tupaia chinensis*	105-146	0.005
I-TASSER			
Internalin K **(2)**	*Listeria monocytogenes*		1.46, 1.43
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		1.21
Antigen MTB48, Mycobacterial protein	*Mycobacterium smegmatis*		1.13
Tropomyosin	*Oryctolagus cuniculus*		1.12

DNA stabilization protein	*Salmonella phage*		1.09
Telomerase-binding protein EST1A	*Homo sapiens*		1.04
Hexon protein	*Human adenovirus 5*		1.01
Human T-cell leukemia virus type II matrix protien	*Human T-lymphotropic virus 2*		1.00
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		0.714
Formin-binding protein 1	*Homo sapiens*		0.632
Phospholipase C beta	*Meleagris gallopavo*		0.632
Cdc42-interacting protein 4	*Homo sapiens*		0.628
Sensor protein torS	*Escherichia coli*		0.611
Tyrosine-protein kinase Fes/Fps	*Homo sapiens*		0.608
GEM-interacting protein	*Homo sapiens*		0.607
SH3-containing GRB2-like protein 2	*Homo sapiens*		0.605
Brain-specific angiogenesis inhibitor 1-associated protein 2-like protein 2	*Mus musculus*		0.604
Predict Protein			
Protein binding		63, 66-72, 74-77	
Secreted			
Atome2			
IQ motif and SEC7 domain-containing protein 1	*Homo sapiens*		80.62
Ubiquitin carboxyl-terminal hydrolase 28	*Homo sapiens*		71.92
Rep (DNA-binding domain)	*Escherichia coli*		66.15
S-adenosylmethionine:tRNA ribosyltransferase-isomerase	*Thermotoga maritima*		65.28
S-adenosylmethionine:tRNA ribosyltransferase-isomerase	*Bacillus subtilis*		54.31
Protein phosphatase 1 regulatory subunit 12A	*Homo sapiens*		50.87
TRPM7 channel	*Rattus norvegicus*		44.52
Nucleoprotein	*Andes virus*		42.17
Adhesin yadA	*Yersinia enterocolitica*		40.69
Activating signal cointegrator 1 complex subunit 2	*Homo sapiens*		36.50
Bacterioferritin	*Pseudomonas aeruginosa*		35.14
Fibroin-modulator binding protine 1	*Bombyx mori*		31.12
Zn(II)-responsive regulator of zntA	*Escherichia coli*		24.78

Estrogen-related receptor gamma	*Homo sapiens*		7.58

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXVII. *Utterbackia imbecillis* H-*ORF* sequences 7 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		71-74	99.28
TIGR04294 prepilin-type processing-associated H-X9-DG domain		13-15	99.16
TIGR01167 LPXTG cell wall anchor domain		49-64	98.89
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		43-45	97.53
TIGR03501 GlyGly-CTERM domain		71-79	96.97
TIGR00756 pentatricopeptide repeat domain		87-98	94.00
BLIstered cuticle family member (bli-2)	*Caenorhabditis elegans*	34-92	94.88
CG12522-PA	*Drosophila melanogaster*	130-198	94.68
F57B1.3	*Caenorhabditis elegans*	29-99	94.44
DumPY: shorter than wild-type family member (dpy-10)	*Caenorhabditis elegans*	23-88	94.43
CG12522-PA	*Drosophila melanogaster*	134-199	94.43
T10E10.5	*Caenorhabditis elegans*	34-96	94.01
DumPY: shorter than wild-type family member (dpy-3)	*Caenorhabditis elegans*	34-102	93.68
DumPY: shorter than wild-type family member (dpy-14)	*Caenorhabditis elegans*	42-89	93.55
Y54E10BL.2	*Caenorhabditis elegans*	44-89	93.36
COLlagen family member (col-102)	*Caenorhabditis elegans*	47-96	92.99
COLlagen family member (col-166)	*Caenorhabditis elegans*	33-96	92.93
SQuaT family member (sqt-2)	*Caenorhabditis elegans*	34-89	92.72
F33A8.9	*Caenorhabditis elegans*	32-96	92.68
F54B11.1	*Caenorhabditis elegans*	33-92	92.25

COLlagen family member (col-101)	*Caenorhabditis elegans*	38-89	92.19
F11G11.11	*Caenorhabditis elegans*	38-93	92.18
COLlagen family member (col-69)	*Caenorhabditis elegans*	32-99	92.09
Armadillo repeat containing, X-linked 4	*Homo sapiens*	102-225	92.06
F55C10.2	*Caenorhabditis elegans*	30-98	92.04
COLlagen family member (col-110)	*Caenorhabditis elegans*	43-99	91.98
T10E10.2	*Caenorhabditis elegans*	34-96	91.71
COLlagen family member (col-2)	*Caenorhabditis elegans*	42-92	91.71
C44C10.1	*Caenorhabditis elegans*	34-96	91.67
COLlagen family member (col-165)	*Caenorhabditis elegans*	34-98	91.35
C12D8.8	*Caenorhabditis elegans*	41-90	91.17
DumPY: shorter than wild-type family member (dpy-9)	*Caenorhabditis elegans*	43-93	91.04
ROLler: helically twisted, animals roll when moving family member (rol-8)	*Caenorhabditis elegans*	44-96	90.69
COLlagen family member (col-14)	*Caenorhabditis elegans*	18-89	90.52
COLlagen family member (col-162)	*Caenorhabditis elegans*	46-90	90.35
C30F2.1	*Caenorhabditis elegans*	45-98	90.17
T11F9.9	*Caenorhabditis elegans*	42-98	90.04
W08D2.6	*Caenorhabditis elegans*	40-91	89.95
COLlagen family member (col-145)	*Caenorhabditis elegans*	46-90	89.58
COLlagen family member (col-36)	*Caenorhabditis elegans*	38-91	89.54
COLlagen family member (col-75)	*Caenorhabditis elegans*	43-89	89.53
Y41C4A.19	*Caenorhabditis elegans*	30-89	89.29
T05A1.2	*Caenorhabditis elegans*	43-89	89.12
F56D5.1	*Caenorhabditis elegans*	33-89	88.98
DumPY: shorter than wild-type family member (dpy-2)	*Caenorhabditis elegans*	26-89	88.62
COLlagen family member (col-43)	*Caenorhabditis elegans*	29-93	88.54
F15H10.1	*Caenorhabditis elegans*	35-92	88.45
F15A2.1	*Caenorhabditis elegans*	41-90	88.29
SQuaT family member (sqt-1)	*Caenorhabditis elegans*	43-89	88.28
COLlagen family member (col-173)	*Caenorhabditis elegans*	43-89	88.21
Axon STeering defect family member (ast-1)	*Caenorhabditis elegans*	103-189	88.10
COLlagen family member (col-114)	*Caenorhabditis elegans*	33-106	87.88
G protein-coupled receptor 152	*Homo sapiens*	2-203	87.87
COLlagen family member (col-120)	*Caenorhabditis elegans*	38-89	87.74
COLlagen family member (col-176)	*Caenorhabditis elegans*	41-89	87.59

Armadillo repeat containing, X-linked 4	*Homo sapiens*	105-209	87.27
F11G11.12	*Caenorhabditis elegans*	35-89	87.16
COLlagen family member (col-183)	*Caenorhabditis elegans*	33-89	87.09
Y41C4A.16	*Caenorhabditis elegans*	30-96	87.00
F08G5.4	*Caenorhabditis elegans*	46-89	86.97
DumPY: shorter than wild-type family member (dpy-8)	*Caenorhabditis elegans*	38-106	86.89
T10E10.1	*Caenorhabditis elegans*	34-98	86.85
COLlagen family member (col-137)	*Caenorhabditis elegans*	42-89	86.42
Reticulon 1 isoform A	*Homo sapiens*	10-95	86.29
Protein involved in bud-site selection	*Saccharomyces cerevisiae*	47-88	85.95
COLlagen family member (col-3)	*Caenorhabditis elegans*	44-93	85.46
COLlagen family member (col-117)	*Caenorhabditis elegans*	44-93	85.46
F32G8.5	*Caenorhabditis elegans*	45-91	85.42
F52F12.2	*Caenorhabditis elegans*	46-99	85.33
AC3.6	*Caenorhabditis elegans*	33-93	85.31
C29F4.1	*Caenorhabditis elegans*	44-96	85.17
T10E10.7	*Caenorhabditis elegans*	23-96	85.12
D2023.7	*Caenorhabditis elegans*	47-93	85.11
CG33203-PC	*Drosophila melanogaster*	2-84	85.07
F14F7.1	*Caenorhabditis elegans*	32-89	85.01
COLlagen family member (col-166)	*Caenorhabditis elegans*	11-96	84.99
COLlagen family member (col-65)	*Caenorhabditis elegans*	37-89	84.85
Lysosomal associated transmembrane protein 4 beta	*Homo sapiens*	12-82	84.84
COLlagen family member (col-45)	*Caenorhabditis elegans*	33-103	84.84
COLlagen family member (col-34)	*Caenorhabditis elegans*	41-92	84.75
F15H10.2	*Caenorhabditis elegans*	35-92	84.65
ROLler: helically twisted, animals roll when moving family member (rol-6)	*Caenorhabditis elegans*	47-89	84.62
COLlagen family member (col-50)	*Caenorhabditis elegans*	43-89	84.57
DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	44-96	84.57
COLlagen family member (col-77)	*Caenorhabditis elegans*	38-93	84.50
ROLler: helically twisted, animals roll when moving family member (rol-8)	*Caenorhabditis elegans*	43-89	84.43
COLlagen family member (col-14)	*Caenorhabditis elegans*	15-82	84.38
F54D1.3	*Caenorhabditis elegans*	34-89	84.34
COLlagen family member (col-93)	*Caenorhabditis elegans*	30-89	84.28

F12F6.9	*Caenorhabditis elegans*	43-89	84.27
Y38C1BA.3	*Caenorhabditis elegans*	44-89	84.23
F54C9.4	*Caenorhabditis elegans*	37-89	84.22
T06E4.4	*Caenorhabditis elegans*	46-89	84.12
BLASTP/PSIBLAST			
Ribonuclease E; Reviewed		94-226	1.13e-06
Ehrlichia tandem repeat (Ehrlichia_rpt)		98-226	3.34e-03
Bv80/Bb-1, partial **(2)**	*Babesia bovis*	98-227	5e-29, 1e-26
PSIBLAST			
Protein B602L, partial	*Columba livia*	106-225	3e-23
Bv80, partial	*Babesia bovis*	105-227	2e-21
Bv80/Bb-1, partial	*Babesia bovis*	101-227	4e-21
Bv80/Bb-1, partial	*Babesia bovis*	101-217	1e-20
85 kDa protein **(2)**	*Babesia bovis*	106-225	1e-20, 4e-20
Bv80, partial	*Babesia bovis*	102-217	3e-20
B602L, partial **(2)**	*African swine fever virus*	82-226	4e-20, 8e-14
Bv80, partial	*Babesia bovis*	105-225	2e-19
9RL	*African swine fever virus*	98-226	1e-18
9RL protein	*African swine fever virus*	87-225	1e-18
9RL	*African swine fever virus*	96-226	2e-18
B602L protein	*African swine fever virus*	88-226	4e-18
Bv80, partial	*Babesia bovis*	105-213	9e-18
B602L, partial	*African swine fever virus*	82-215	2e-17
B602L protein	*African swine fever virus*	87-222	3e-17
9RL	*African swine fever virus*	103-225	3e-16
B602L, partial	*African swine fever virus*	82-211	4e-16
B602L protein	*African swine fever virus*	88-223	7e-16
B602L **(2)**	*African swine fever virus*	87-215	1e-15, 4e-15
B602L protein	*African swine fever virus*	111-219	2e-15
B602L, partial	*African swine fever virus*	82-203	1e-14
S-layer protein precursor	*Bacillus thuringiensis*	102-190	1e-14
pB602L **(2)**	*African swine fever virus*	82-225	2e-14, 7e-

			14
Cell surface protein **(3)**	*Bacillus thuringiensis*	102-190	2e-14-4e-14
Bv80/Bb-1, partial	*Babesia bovis*	107-178	8e-14
B602L protein	*African swine fever virus*	107-219	8e-14
BV80 merozoite protein	*Babesia bovis*	109-201	1e-13
9RL	*African swine fever virus*	111-219	2e-13
B602L protein **(2)**	*African swine fever virus*	87-203	5e-13, 7e-13
B602L, partial **(4)**	*African swine fever virus*	82-195	5e-13, 4e-11
85 kDa protein	*Babesia bovis*	99-177	5e-13
Type IV secretion protein Rhs, partial	*Nocardioides sp. URHA0020*	102-195	1e-12
9RL	*African swine fever virus*	103-203	2e-12
Glutamate/valine-rich protein	*Natronorubrum sulfidifaciens*	54-223	2e-12
9RL protein **(2)**	*African swine fever virus*	87-203	2e-12, 4e-12
Liver stage antigen 3	*Plasmodium falciparum*	102-227	3e-12
B602L **(2)**	*African swine fever virus*	82-207	3e-12, 1e-08
B602L protein **(8)**	*African swine fever virus*	103-203	3e-12-2e-11
B602L protein **(5)**	*African swine fever virus*	87-188	4e-12-2e-09
B602L protein, partial **(3)**	*African swine fever virus*	87-195	4e-12-4e-11
9RL protein	*African swine fever virus*	87-179	5e-12
pB602L, partial	*African swine fever virus*	87-209	1e-11
Central variable region protein	*African swine fever virus*	82-188	2e-11
B602L, partial	*African swine fever virus*	82-188	5e-11
B602L protein	*African swine fever virus*	87-191	6e-11
GH24581	*Drosophila grimshawi*	99-212	6e-11
B602L, partial	*African swine fever virus*	82-179	1e-10
Cellulosomal scaffoldin anchoring protein	*Trichomonas vaginalis*	98-226	2e-10
9RL protein, partial	*African swine fever virus*	103-218	3e-10
Transcription factor IIIB 50 kDa subunit	*Xenopus tropicalis*	48-226	3e-10

pB602L	*African swine fever virus*	82-215	4e-10
B602L protein	*African swine fever virus*	107-226	2e-09
pB602L	*African swine fever virus*	82-219	2e-09
9RL	*African swine fever virus*	96-179	2e-09
Chitinase	*Deefgea rivuli*	104-215	8e-09
9RL protein, partial	*African swine fever virus*	107-223	2e-08
pB602L	*African swine fever virus*	82-203	3e-08
PT repeat family protein	*Aspergillus fumigatus*	102-174	6e-08
Glutamate/valine-rich protein	*Halosarcina pallida*	108-227	8e-08
Cellulosomal scaffoldin anchoring protein C	*Trichomonas vaginalis*	98-226	2e-07
PT repeat family protein	*Neosartorya fischeri*	100-210	2e-07
Liver stage antigen 3	*Plasmodium falciparum*	92-225	4e-07, 3e-05
LdOrf-129 peptide	*Lymantria dispar multiple nucleopolyhedrovirus*	149-222	6e-07
Liver stage antigen 3	*Plasmodium falciparum*	92-227	1e-06
9RL protein, partial **(2)**	*African swine fever virus*	131-223	1e-06
Pullulanase, type I	*Lachnospiraceae bacterium*	98-195	1e-06
Liver stage antigen 3 precursor	*Plasmodium knowlesi*	90-227	2e-06
B602L protein	*African swine fever virus*	131-223	2e-06
CG3108	*Drosophila melanogaster*	97-222	2e-06
pB602L	*African swine fever virus*	82-185	5e-06
ORF-132 protein	*Lymantria dispar multiple nucleopolyhedrovirus*	134-227	7e-06
Involucrin repeat protein	*Ophiostoma piceae*	114-226	7e-06
BA71V-B602L (9RL)	*African swine fever virus*	82-187	9e-06
Snaclec 3	*Toxocara canis*	97-225	1e-05
LPXTG-motif cell wall anchor domain protein	*Nannochloropsis gaditana*	101-223	2e-05
Central variable region protein	*African swine fever virus*	82-178	2e-05
Pathway-specific nitrogen regulator	*Metarhizium guizhouense*	98-186	2e-05
9RL protein **(2)**	*African swine fever virus*	82-178	2e-05, 3e-05
Liver stage antigen-3	*Plasmodium falciparum*	88-227	4e-05
Glutamate/valine-rich protein	*Natrinema pallidum*	33-227	5e-05
Liver stage antigen 3	*Plasmodium falciparum*	102-195	7e-05
Liver stage antigen 3	*Plasmodium falciparum*	103-227	7e-05

9RL protein **(2)**	*African swine fever virus*	139-223	8e-05, 9e-05
Liver stage antigen 3	*Plasmodium falciparum*	102-227	8e-05
Liver stage antigen 3	*Plasmodium falciparum*	99-227	9e-05
Glutamate/valine-rich protein	*Natrinema sp. J7-2*	33-227	1e-04
B602L protein	*African swine fever virus*	139-223	1e-04
Cell division protein, partial	*Streptococcus sanguinis*	113-225	1e-04
Glutamate/valine-rich protein	*Natronorubrum tibetense*	45-187	2e-04
Liver stage antigen 3	*Ectocarpus siliculosus*	93-219	2e-04
Glutamate/valine-rich protein	*Natrinema gari*	33-227	2e-04
von Willebrand factor type A	*Streptomyces fulvissimus*	109-187	3e-04
Liver stage antigen 3	*Plasmodium falciparum*	102-225	4e-04, 0.001
Liver stage antigen-3	*Plasmodium falciparum*	103-225	4e-04
Glutamate/valine-rich protein	*Natrinema altunense*	102-223	0.001
Cellulosomal scaffoldin anchoring protein	*Trichomonas vaginalis*	99-226	0.002
Liver stage antigen 3	*Plasmodium falciparum*	107-227	0.004
I-TASSER			
Survival motor neuron protein **(4)**	*Homo sapiens*		1.12-1.01
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		2.56
Myc box dependent interacting protein 1 **(2)**	*Homo sapiens*		1.13, 1.94
Accumulation associated protein	*Staphylococcus epidermidis*		1.56
60S ribosomal protein L28	*Saccharomyces cerevisiae*		2.32
Survival motor neuron protein **(4)**	*Homo sapiens*		1.12-1.01
Predict Protein			
Protein binding		8, 66	
Polynucleotide binding		27, 32, 34-35	
Cytoplasm			
Opioid growth factor receptor **(4)**	*Homo sapiens*		2e-11-2e-08
Proteoglycan 4 **(10)**	*Mus musculus*		2e-08-0.92
Cell surface glycoprotein 1 **(17)**	*Clostridium thermocellum*		0.001-0.010
Atome2			

TuSp1	*Nephila antipodiana*		78.07
80 kDa MCM3-associated protein	*Homo sapiens*		55.44
12 kDa heat shock protein	*Saccharomyces cerevisiae*		45.97
Interferon alpha-inducible protein 27-like protein 1	*Homo sapiens*		24.35
ICP47	*Herpes simplex virus*		1.72

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXVIII. *Margaritifera margaritifera* H-*ORF* sequence 1 function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		5-7	99.23
TIGR04294 prepilin-type processing-associated H-X9-DG domain		5-8	99.16
TIGR01167 LPXTG cell wall anchor domain		43-59	98.88
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		60-67	97.66
TIGR03501 GlyGly-CTERM domain		47-59	96.99
TIGR00756 pentatricopeptide repeat domain		23-27	93.25
T24B1.1	*Caenorhabditis elegans*	21-63	91.85
Occlusion-derived virus envelope protein ODV-E18		33-73	74.97
d.24.1 Pili subunits (54523) SCOP seed sequence: d1oqwa_		42-61	74.24
d.24.1 Pili subunits (54523) SCOP seed sequence: d2pila_		42-61	74.20
Serine protease inhibitor		48-107	71.24
Chitin synthesis regulation, resistance to Congo red		43-61	70.62
Activator of basal transcription 1	*Homo sapiens*	5-33	69.40
Occlusion-derived virus envelope protein ODV-E18		34-73	68.30

CG17785-PA	*Drosophila melanogaster*	16-65	68.04
General secretion pathway protein H	*Beggiatoa sp. PS*	42-61	67.36
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-90)	*Caenorhabditis elegans*	45-64	63.72
Secreted protein	*Beggiatoa sp. PS*	42-61	61.88
CG32708-PA	*Drosophila melanogaster*	5-33	61.30
Activator of basal transcription	*Mus musculus*	5-33	59.95
COLlagen family member (col-93)	*Caenorhabditis elegans*	29-67	57.32
Alpha defensin		50-61	57.29
COLlagen family member (col-34)	*Caenorhabditis elegans*	30-67	56.89
C17H11.6c	*Caenorhabditis elegans*	20-70	56.74
Cytochrome c550	*Bacillus subtilis*	38-72	56.20
RCR		43-61	52.93
ComB		18-61	52.89
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-2)	*Caenorhabditis elegans*	45-66	52.79
COLlagen family member (col-91)	*Caenorhabditis elegans*	42-67	51.23
UCP036704		29-35	51.11
TATA-binding protein binding **(2)**	*Arabidopsis thaliana*	5-33	50.96
General secretion pathway protein J	*Yersinia pestis*	42-61	50.70
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-78)	*Caenorhabditis elegans*	45-66	50.59
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-79)	*Caenorhabditis elegans*	45-66	49.73
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-76)	*Caenorhabditis elegans*	45-66	49.73
Essential nucleolar protein involved in pre-18S rRNA processing	*Saccharomyces cerevisiae*	5-33	49.71
CG6999-PA	*Drosophila melanogaster*	5-33	49.32
T-cell receptor-associates transmembrane adapter 1		42-59	49.10
Thymidine kinase **(2)**	*Herpes virus*	23-34	48.70
K08F4.5	*Caenorhabditis elegans*	42-61	48.36
Essential cell division protein	*Escherichia coli*	36-85	48.20
GRP: Glycine rich protein family		46-64	47.31
FAST kinase-like protein, subdomain 1		5-54	46.87
COLlagen family member (col-94)	*Caenorhabditis elegans*	29-67	46.72

TIGR03544 DivIVA domain	*Bacillus subtilis*	20-35	51.87
COLlagen family member (col-165)	*Caenorhabditis elegans*	29-67	45.84
Activated in Blocked Unfolded protein response family member (abu-7)	*Caenorhabditis elegans*	45-66	45.83
PulG Type II secretory pathway, pseudopilin PulG		42-61	45.03
COLlagen family member (col-139)	*Caenorhabditis elegans*	30-67	44.54
F27E5.3	*Caenorhabditis elegans*	42-61	44.44
Light-harvesting complex subunits	*Rhodospirillum rubrum*	39-63	43.91
General secretion pathway protein G	*Beggiatoa sp. PS*	42-61	43.24
COLlagen family member (col-92)	*Caenorhabditis elegans*	29-67	43.10
General secretion pathway protein H	*Beggiatoa sp. PS*	42-61	43.05
CG32706-PA	*Drosophila melanogaster*	5-33	42.51
ABC transporter, permease protein	*Methanosarcina mazei*	11-64	42.42
COLlagen family member (col-102)	*Caenorhabditis elegans*	42-67	41.89
Light-harvesting complex subunits	*Rhodospirillum rubrum*	39-63	41.47
COLlagen family member (col-114)	*Caenorhabditis elegans*	29-67	41.33
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-91)	*Caenorhabditis elegans*	45-66	41.32
PulG: Type II secretory pathway pseudopilin		42-61	41.30
F420-nonreducing hydrogenase II, subunit cytochrome B	*Methanosarcina mazei*	33-61	41.29
EGF-like-domain, multiple 9 **(2)**	*Homo sapiens*	45-62	40.73
Y45G12B.2a	*Caenorhabditis elegans*	37-62	40.38
I-TASSER			
Tropomyosin	*Oryctolagus cuniculus*		1.12
Nck-associated protein 1	*Homo sapiens*		0.511
Predict Protein			
Protein binding		16, 19, 27-31, 34-37, 60-61, 64, 92	
Secreted			
Atome2			
39 kDa initiator binding protein	*Trichomonas vaginalis*		67.83
Octamer-binding transcription factor 1	*Homo sapiens*		59.87
BA3-type cytochrome-c oxidase	*Thermus thermophilus*		55.07

Electron transfer flavoprotein-ubiquinone oxidoreductase	*Sus scrofa*		50.03
Degenerin mec-4	*Caenorhabditis elegans*		48.90
Cytochrome b-c1 complex subunit 1, mitochondrial	*Saccharomyces cerevisiae*		28.76
Integrin alpha-IIb	*Homo sapiens*		28.20
Fimbrial protein	*Neisseria gonorrhoeae*		23.71
Fimbrial protein	*Pseudomonas aeruginosa*		21.80
Fimbrial protein	*Dichelobacter nodosus*		21.07
Cycloviolacin O14	*Viola odorata*		17.54
Histone peptide	*Homo sapiens*		15.68
Cytochrome c oxidase subunit 1	*Thermus thermophilus*		13.57

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XXXIX. *Margaritifera margaritifera* H-*ORF* sequences 2 & 4 function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		5-7	99.23
TIGR04294 prepilin-type processing-associated H-X9-DG domain		5-8	99.16
TIGR01167 LPXTG cell wall anchor domain		43-60	98.93
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		61-67	97.50
TIGR03501 GlyGly-CTERM domain		47-59	96.99
TIGR00756 pentatricopeptide repeat domain		23-27	93.25
T24B1.1	*Caenorhabditis elegans*	21-63	92.10
d.24.1 Pili subunits (54523) SCOP seed sequence: d2pila_		42-61	75.70
RCR		43-61	75.43

d.24.1 Pili subunits (54523) SCOP seed sequence: d1oqwa_		42-61	75.42
Occlusion-derived virus envelope protein ODV-E18		33-73	75.30
Activator of basal transcription 1	*Homo sapiens*	5-33	69.31
Activated in Blocked Unfolded protein response family member (abu-1)	*Caenorhabditis elegans*	45-64	69.18
Occlusion-derived virus envelope protein ODV-E18		34-73	68.67
Serine protease inhibitor		48-107	68.30
General secretion pathway protein H	*Beggiatoa sp. PS*	42-61	65.38
Secreted protein	*Beggiatoa sp. PS*	42-61	64.26
CG17785-PA	*Drosophila melanogaster*	16-65	64.05
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-90)	*Caenorhabditis elegans*	45-64	62.89
CG32708-PA	*Drosophila melanogaster*	5-33	61.19
Activator of basal transcription	*Mus musculus*	5-33	59.86
ComB		18-61	59.66
C17H11.6c	*Caenorhabditis elegans*	20-70	59.31
RCR		43-61	58.19
COLlagen family member (col-34)	*Caenorhabditis elegans*	30-67	56.67
Cytochrome c550	*Bacillus subtilis*	38-72	56.62
COLlagen family member (col-93)	*Caenorhabditis elegans*	29-67	54.67
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-2)	*Caenorhabditis elegans*	45-66	52.52
General secretion pathway protein J	*Yersinia pestis*	42-61	51.73
UCP036704		29-35	51.14
TATA-binding protein binding	*Arabidopsis thaliana*	5-33	50.88
TATA-binding protein binding	*Arabidopsis thaliana*	5-33	50.88
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-78)	*Caenorhabditis elegans*	45-66	50.32
T-cell receptor-associated transmembrane adapter 1		42-59	49.77
Essential nucleolar protein involved in pre-18S rRNA processing	*Saccharomyces cerevisiae*	5-33	49.60
Glycine rich protein family		46-64	49.40
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-79)	*Caenorhabditis elegans*	45-66	49.22
Prion-like-(Q/N-rich)-domain-bearing protein family	*Caenorhabditis elegans*	45-66	49.22

member (pqn-76)			
CG6999-PA	*Drosophila melanogaster*	5-33	49.21
Thymidine kinase	*Herpes virus*	23-34	48.81
K08F4.5	*Caenorhabditis elegans*	42-61	48.72
Thymidine kinase	*Herpes virus*	23-34	48.23
COLlagen family member (col-91)	*Caenorhabditis elegans*	42-67	47.33
PulG Type II secretory pathway, pseudopilin PulG		42-61	47.33
Membrane spanning protein in TonB-ExbB-ExbD complex	*Escherichia coli*	41-85	46.99
FAST kinase-like protein, subdomain 1		5-54	46.88
Alpha defensin		50-61	46.56
TIGR03544 DivIVA domain		20-35	51.80
Activated in Blocked Unfolded protein response family member (abu-7)	*Caenorhabditis elegans*	45-66	45.52
F27E5.3	*Caenorhabditis elegans*	42-61	44.20
COLlagen family member (col-94)	*Caenorhabditis elegans*	29-67	43.99
COLlagen family member (col-165)	*Caenorhabditis elegans*	29-67	43.29
General secretion pathway protein G	*Beggiatoa sp. PS*	42-61	43.29
ABC transporter, permease protein	*Methanosarcina mazei*	11-64	43.09
CG32706-PA	*Drosophila melanogaster*	5-33	42.39
COLlagen family member (col-139)	*Caenorhabditis elegans*	30-67	41.82
PulG: Type II secretory pathway pseudopilin		42-61	41.79
Essential cell division protein	*Escherichia coli*	36-85	41.62
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-91)	*Caenorhabditis elegans*	45-66	41.35
General secretion pathway protein H	*Beggiatoa sp. PS*	42-61	41.33
RCR		42-62	41.12
Integral membrane protein	*Streptomyces coelicolor*	15-66	41.11
F420-nonreducing hydrogenase II, subunit cytochrome B	*Methanosarcina mazei*	33-61	41.03
Type II secretion system protein I		39-63	40.83
DevC protein	*Nostoc punctiforme*	14-64	40.55
I-TASSER			
Tropomyosin	*Oryctolagus cuniculus*		1.07
Predict Protein			
Protein binding		1-2, 16,	

		27-31, 34-38, 60, 92, 100-102	
Mitochondrion membrane			
Atome2			
39 kDa initiator binding protein (C-domain, residues 127-341)	*Trichomonas vaginalis*		84.16
Electron transfer flavoprotein-ubiquinone oxidoreductase	*Sus scrofa*		63.11
Photosystem Q(B) protein	*Thermosynechococcus vulcanus*		39.25
Photosystem Q(B) protein	*Thermosynechococcus elongatus*		37.01
Apocytochrome f	*Chlamydomonas reinhardtii*		32.77
Integrin alpha-IIb (transmembrane and cytoplasmic domains, residues 991-1039)	*Homo sapiens*		32.70
Cytochrome b-c1 complex subunit 1, mitochondrial	*Saccharomyces cerevisiae*		32.68
Cytochrome b6 **(2)**	*Mastigocladus laminosus*		26.41, 26.19
Fimbrial protein	*Neisseria gonorrhoeae*		26.33
Fimbrial protein	*Pseudomonas aeruginosa*		24.49
Fimbrial protein	*Dichelobacter nodosus*		23.78
Cycloviolacin O14	*Viola odorata*		23.59
Cytochrome c oxidase subunit 1	*Thermus thermophilus*		14.67

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XL. *Margaritifera margaritifera* H-*ORF* sequence 3 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		5-7	99.19
TIGR04294 prepilin-type processing-associated H-X9-DG domain		5-8	99.14
TIGR01167 LPXTG cell wall anchor domain		43-60	98.94
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		61-67	97.43
TIGR03501 GlyGly-CTERM domain		47-60	96.95
TIGR00756 pentatricopeptide repeat domain		23-27	93.37
T24B1.1	*Caenorhabditis elegans*	21-63	91.45
d.24.1 Pili subunits (54523) SCOP seed sequence: d2pila_		42-61	79.09
d.24.1 Pili subunits (54523) SCOP seed sequence: d1oqwa_		42-61	78.95
RCR		43-61	77.34
Occlusion-derived virus envelope protein ODV-E18		33-73	77.10
Activated in Blocked Unfolded protein response family member (abu-1)	*Caenorhabditis elegans*	45-64	70.87
Occlusion-derived virus envelope protein ODV-E18		34-66	70.71
General secretion pathway protein H	*Beggiatoa sp. PS*	42-61	69.93
Secreted protein	*Beggiatoa sp. PS*	42-61	67.77
Activator of basal transcription 1	*Homo sapiens*	5-33	66.74
C17H11.6c	*Caenorhabditis elegans*	20-70	63.78
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-90)	*Caenorhabditis elegans*	45-64	63.28
CG17785-PA	*Drosophila melanogaster*	16-65	61.12
RCR		43-61	60.93
COLlagen family member (col-34)	*Caenorhabditis elegans*	30-67	58.47
CG32708-PA	*Drosophila melanogaster*	5-33	58.17
Activator of basal transcription	*Mus musculus*	5-33	57.26
Serine protease inhibitor		48-116	56.16
COLlagen family member (col-93)	*Caenorhabditis elegans*	29-67	55.87

General secretion pathway protein J	*Yersinia pestis*	42-61	55.83
Cytochrome c550	*Bacillus subtilis*	38-72	54.89
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-2)	*Caenorhabditis elegans*	45-66	54.43
T-cell receptor-associated transmembrane adapter 1		42-59	53.90
Membrane spanning protein in TonB-ExbB-ExbD complex	*Escherichia coli*	41-85	52.51
COLlagen family member (col-91)	*Caenorhabditis elegans*	42-67	52.18
PulG Type II secretory pathway, pseudopilin PulG		42-61	51.91
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-79)	*Caenorhabditis elegans*	45-66	51.81
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-76)	*Caenorhabditis elegans*	45-66	51.81
Glycine rich protein family		46-64	51.77
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-78)	*Caenorhabditis elegans*	45-66	51.24
ComB		18-61	49.37
K08F4.5	*Caenorhabditis elegans*	42-61	48.67
Alpha defensin		50-61	48.02
TATA-binding protein binding	*Arabidopsis thaliana*	5-33	47.96
TATA-binding protein binding	*Arabidopsis thaliana*	5-33	47.96
FAST kinase-like protein, subdomain 1		5-54	47.41
Essential nucleolar protein involved in pre-18S rRNA processing	*Saccharomyces cerevisiae*	5-33	46.17
CG6999-PA	*Drosophila melanogaster*	5-33	46.16
Activated in Blocked Unfolded protein response family member (abu-7)	*Caenorhabditis elegans*	45-60	45.70
F27E5.3	*Caenorhabditis elegans*	42-61	45.70
COLlagen family member (col-94)	*Caenorhabditis elegans*	29-67	45.33
Type II secretion system protein I		39-63	44.95
F18A12.1	*Caenorhabditis elegans*	42-99	44.63
General secretion pathway protein H	*Beggiatoa sp. PS*	32-61	44.03
COLlagen family member (col-139)	*Caenorhabditis elegans*	30-67	43.90
COLlagen family member (col-102)	*Caenorhabditis elegans*	42-67	43.79
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-91)	*Caenorhabditis elegans*	45-68	43.37

UCP036704		29-35	42.69
RCR		42-62	42.53
COLlagen family member (col-165)	*Caenorhabditis elegans*	25-67	41.85
F38A3.1	*Caenorhabditis elegans*	25-67	41.82
ABC transporter, permease protein	*Methanosarcina mazei Go1*	11-64	41.51
EGF-like-domain, multiple 9 **(2)**	*Homo sapiens*	45-62	41.43
PulG: Type II secretory pathway pseudopilin		42-61	41.24
Essential cell division protein	*Escherichia coli K12*	33-85	41.21
Y54E10BL.2	*Caenorhabditis elegans*	44-67	41.18
COLlagen family member (col-133)	*Caenorhabditis elegans*	29-67	41.00
TIGR03544 DivIVA domain		20-35	46.94
F420-nonreducing hydrogenase II, subunit cytochrome B	*Methanosarcina mazei Go1*	33-61	40.67
COLlagen family member (col-92)	*Caenorhabditis elegans*	29-67	40.64
PSIBLAST			
BnaC08g34590D	*Brassica napus*	43-134	2e-04, 3e-04
I-TASSER			
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		1.63
Predict Protein			
Protein binding		1, 11-13, 27-31, 35, 37-39, 91-92, 98-104	
Secreted			
Atome2			
39 kDa initiator binding protein	*Trichomonas vaginalis*		85.93
Electron transfer flavoprotein-ubiquinone oxidoreductase	*Sus scrofa*		64.08
Degenerin mec-4	*Caenorhabditis elegans*		51.62
Adenovirus fiber	*Human adenovirus 2*		34.11
Receptor tyrosine-protein kinase erbB-3	*Homo sapiens*		30.21
Fiber protein 2	*Human adenovirus 41*		29.74
Cytochrome b6 **(4)**	*Mastigocladus laminosus*		28.52-

			20.59
Nitrate/TMAO reductases, membrane-bound tetraheme cytochrome c subunit	*Desulfovibrio alaskensis*		26.62
Fimbrial protein	*Neisseria gonorrhoeae*		25.11
Fimbrial protein	*Pseudomonas aeruginosa*		22.24
Fimbrial protein	*Dichelobacter nodosus*		20.69

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XLI. *Toxolasma lividus* H-*ORF* function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR01167 LPXTG cell wall anchor domain		24-41	99.01
TIGR04294 prepilin-type processing-associated H-X9-DG domain		134-135	98.98
TIGR03304 outer membrane insertion C-terminal signal		40-44	98.46
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		189-194	97.44
TIGR03501 GlyGly-CTERM domain		33-44	96.69
TIGR00756 pentatricopeptide repeat domain		17-23	93.27
CG17248-PA, isoform A	*Drosophila melanogaster*	16-81	91.85
CG17248-PC, isoform C	*Drosophila melanogaster*	16-81	91.85
CG17248-PE, isoform E	*Drosophila melanogaster*	16-81	90.04
CG17248-PB, isoform B	*Drosophila melanogaster*	16-81	90.04
G-protein-linked Acetylcholine Receptor family member (gar-1)	*Caenorhabditis elegans*	6-52	84.25
sensor protein	*Nostoc punctiforme*	2-43	81.29
cAMP responsive element binding protein 3-like 2	*Mus musculus*	8-50	80.14
CG33517-PB, isoform B	*Drosophila melanogaster*	6-52	78.13
Vesicle-associated membrane protein 2 (synaptobrevin 2)	*Homo sapiens*	16-43	78.05

CG3856-PC, isoform C	*Drosophila melanogaster*	8-52	77.12
CG3856-PA, isoform A	*Drosophila melanogaster*	8-52	77.12
CG33517-PC, isoform C	*Drosophila melanogaster*	8-52	76.96
Histamine receptor-related G-protein coupled receptor		8-52	76.91
K08F4.5	*Caenorhabditis elegans*	20-43	76.75
ATGLR2.4	*Arabidopsis thaliana*	4-39	76.65
Septation ring formation regulator EzrA	*Staphylococcus aureus*	20-42	76.59
ABC Transporter	*Sulfolobus solfataricus*	9-59	76.52
C-type LECtin family member (clec-39)	*Caenorhabditis elegans*	8-47	76.41
G-protein-linked Acetylcholine Receptor family member (gar-2)	*Caenorhabditis elegans*	3-52	75.76
Syndecan 3	*Mus musculus*	10-42	75.43
G-protein-linked Acetylcholine Receptor family member (gar-2)	*Caenorhabditis elegans*	6-52	74.50
Integral membrane sensor signal transduction histidine kinase	*Nostoc punctiforme*	14-44	74.45
F02E9.7	*Caenorhabditis elegans*	3-49	74.18
Dentin matrix protein 1	*Mus musculus*	32-49	73.94
COLlagen family member (col-77)	*Caenorhabditis elegans*	8-42	73.16
C17H11.6c	*Caenorhabditis elegans*	1-52	73.01
CG4356-PA, isoform A	*Drosophila melanogaster*	6-63	72.19
TonB family protein	*Nostoc punctiforme*	1-45	71.31
Membrane-bound protease FTSH (cell division protein)	*Mycobacterium tuberculosis*	14-40	71.19
DumPY: shorter than wild-type family member (dpy-2)	*Caenorhabditis elegans*	4-42	70.87
Histidine kinase	*Nitrosopumilus maritimus*	13-42	70.83
Y26D4A.6	*Caenorhabditis elegans*	28-49	70.46
CG16720-PB, isoform B	*Drosophila melanogaster*	8-52	69.15
CG16720-PA, isoform A	*Drosophila melanogaster*	8-52	69.15
C49D10.10	*Caenorhabditis elegans*	14-48	68.66
G-protein-linked Acetylcholine Receptor family member (gar-2)	*Caenorhabditis elegans*	8-52	68.38
Phosphonate ABC transporter	*Nostoc punctiforme*	13-37	68.13
T06E4.6	*Caenorhabditis elegans*	19-42	66.32
Syntaxin		8-41	66.31
CG18208-PA	*Drosophila melanogaster*	6-52	65.73
Cuticle protein		30-49	65.60

LCR9	*Arabidopsis thaliana*	19-39	65.51
Cation efflux system protein czcA-1	*Synechococcus sp. CC9311*	17-72	65.27
Extracellular solute-binding protein	*Thermofilum pendens*	14-38	64.33
T06E4.8	*Caenorhabditis elegans*	14-61	64.21
COLlagen family member (col-77)	*Caenorhabditis elegans*	6-40	63.74
Cytochrome c-550	*Nostoc punctiforme*	19-40	63.62
DOPamine receptor family member (dop-3)	*Caenorhabditis elegans*	1-52	63.59
N-terminal TM domain of oligopeptide transport permease C		30-69	63.54
K08F4.5	*Caenorhabditis elegans*	26-47	62.68
Structural constituent of cell wall	*Arabidopsis thaliana*	4-71	62.58
Glycophorin		23-43	62.12
Vesicle-associated membrane protein 8	*Mus musculus*	4-45	62.00
High affinity copper uptake protein 1	*Homo sapiens*	32-43	61.97
Cell-division initiation protein	*Bacillus subtilis*	8-43	61.88
C44C10.1	*Caenorhabditis elegans*	8-40	61.69
COLlagen family member (col-173)	*Caenorhabditis elegans*	5-44	60.97
CG9778-PA	*Drosophila melanogaster*	20-78	60.97
Molybdate-binding protein	*Methanosarcina mazei*	12-43	60.58
Vesicle-associated membrane protein 1 isoform 1	*Homo sapiens*	16-43	60.32
LCR5	*Arabidopsis thaliana*	19-39	60.27
serine protease SplA	*Staphylococcus aureus*	14-33	60.15
RNase_BN		6-44	60.11
Activated in Blocked Unfolded protein response family member (abu-11)	*Caenorhabditis elegans*	32-44	60.05
Sulfate ABC transporter, periplasmic sulfate-binding protein	*Nostoc punctiforme*	5-36	59.68
Cholinergic receptor, muscarinic 3	*Homo sapiens*	8-52	59.63
DOPamine receptor family member (dop-2)	*Caenorhabditis elegans*	6-52	59.47
DOPamine receptor family member (dop-3)	*Caenorhabditis elegans*	10-52	59.43
I-TASSER			
Survival motor neuron protein	*Homo sapiens*		1.97, 1.36
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		2.63, 1.18
Myc box dependent interacting protein 1	*Homo sapiens*		1.39
Accumulation associated protein	*Staphylococcus epidermidis*		1.65

Major vault protein	*Rattus norvegicus*		1.61
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		0.716
Dynamin family protein	*Nostoc punctiforme*		0.606
Tyrosine-protein kinase Fes/Fps	*Homo sapiens*		0.563
Interferon-induced guanylate-binding protein 1	*Homo sapiens*		0.559
BAI1-associated protein 2 isoform 1	*Homo sapiens*		0.546
Phospholipase C beta	*Meleagris gallopavo*		0.544
Brain-specific angiogenesis inhibitor 1-associated protein 2-like protein 2	*Mus musculus*		0.539
TcdA1	*Photorhabdus luminescens*		0.537
Predict Protein			
Protein binding		1-2	
Secreted			
Cell surface protein **(5)**	*Bacillus cereus*		1e-06-5e-05
DNA-directed RNA polymerase II subunit RPB1	*Homo sapiens*		0.30
DNA-directed RNA polymerase II subunit RPB1	*Mus musculus*		0.29
Agglutinin receptor	*Streptococcus gordonii*		0.28
Muscle M-line assembly protein unc-89	*Caenorhabditis elegans*		
Atome2			
Glucose-1-phosphate thymidylyl transferase	*Pseudomonas aeruginosa*		72.22
VPU **(3)**	*Human immunodeficiency virus 1*		33.16-27.03
GDP-mannose mannosyl hydrolase	*Escherichia coli*		30.48

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XLII. *Lasmigona compressa* H-*ORF* sequence 1 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		17-20	99.15
TIGR01167 LPXTG cell wall anchor domain		2-13	98.90
TIGR03304 outer membrane insertion C-terminal signal		23-24	98.77
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		79-92	97.64
TIGR03501 GlyGly-CTERM domain		2-12	96.49
TIGR00756 pentatricopeptide repeat domain		14-34	93.03
C53B4.8	*Caenorhabditis elegans*	75-198	87.53
C53B4.8	*Caenorhabditis elegans*	52-196	86.53
W02B8.6	*Caenorhabditis elegans*	53-198	71.92
W02B8.6	*Caenorhabditis elegans*	76-198	70.34
W02B8.3	*Caenorhabditis elegans*	51-204	56.63
F32A11.7	*Caenorhabditis elegans*	30-198	56.42
TonB family protein	*Nostoc punctiforme*	2-25	50.10
W02B8.3	*Caenorhabditis elegans*	51-198	49.98
RNA binding/nucleic acid binding	*Arabidopsis thaliana*	55-207	46.78
W02B8.4	*Caenorhabditis elegans*	53-198	42.83
Sterol reductase/lamin B receptor		7-36	40.81
Related to CHS7 - control of protein export from the ER (like chitin synthase III)		3-27	40.45
COLlagen family member (col-102)	*Caenorhabditis elegans*	1-35	40.42
Neuropeptide-Like Protein family member (nlp-16)	*Caenorhabditis elegans*	1-17	40.23
RNA binding/nucleic acid binding	*Arabidopsis thaliana*	48-202	39.87
Delta-notch-like EGF repeat-containing transmembrane	*Homo sapiens*	1-101	39.77
W02B8.4	*Caenorhabditis elegans*	76-196	38.59
F19H8.4	*Caenorhabditis elegans*	75-196	35.09
Defensin, beta 104B precursor **(2)**	*Homo sapiens*	1-24	33.93
Chs3p: Chitin synthase III catalytic subunit **(2)**		3-34	33.08, 28.76
F58A4.1	*Caenorhabditis elegans*	2-10	29.62
Y41C4A.19	*Caenorhabditis elegans*	1-35	27.84

Phosphatidylinositol glycan, class B	*Homo sapiens*	2-25	27.40
Subunit X of cytochrome bc1 complex	*Saccharomyces cerevisiae*	1-25	25.83
CTAGE family, member 5 isoform 1	*Homo sapiens*	1-29	25.31
Plasmodium falciparum S-antigen		1-19	24.47
RCR		1-19	24.31
B0379.7	*Caenorhabditis elegans*	2-19	24.10
CG7685-PA	*Drosophila melanogaster*	2-30	23.18
CG8764-PA	*Drosophila melanogaster*	2-25	22.39
Sar8.2 family		1-42	22.36
Subunit 9 of the ubiquinol cytochrome-c reductase complex	*Saccharomyces cerevisiae*	1-25	22.29
Ergosterol biosynthesis ERG4/ERG24 family		8-36	22.13
C35A5.4	*Caenorhabditis elegans*	47-61	22.03
Activated in Blocked Unfolded protein response family member (abu-11)	*Caenorhabditis elegans*	5-29	21.87
T26E3.1	*Caenorhabditis elegans*	1-11	21.62
SRB6		11-30	21.50
SH3 type 3 domain-containing protein	*Nostoc punctiforme*	2-21	21.10
DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	1-30	21.03
Transmembrane protein	*Mycobacterium tuberculosis*	1-31	20.90
Cytochrome b-c1 complex subunit 9	*Saccharomyces cerevisiae*	1-25	20.65
PVC2		1-34	20.55
C04H5.7	*Caenorhabditis elegans*	2-23	20.51
F29B9.9	*Caenorhabditis elegans*	1-25	20.18
Transcriptional regulator, XRE family	*Beggiatoa sp. PS*	10-32	20.10
Ubiquinol-cytochrome c reductase complex 7.2kDa protein isoform a	*Homo sapiens*	2-25	20.01
BLASTP/PSIBLAST			
Viral protein TPX	*Histoplasma capsulatum*	39-201	1e-10
Proteoglycan	*Histoplasma capsulatum*	32-199	6e-10, 5e-09
Histone-lysine N-methyltransferase ATXR3	*Medicago truncatula*	77-199	1e-04
Histone-lysine N-methyltransferase E(z)	*Medicago truncatula*	77-199	1e-04
Adhesin	*Rahnella aquatilis*	43-192	2e-04
BNIP2 motif containing molecule at the carboxyl terminal region 1-like protein	*Camelus ferus*	35-191	3e-04

Chitinase III	*Acanthocheilonema viteae*	35-199	0.001
BLASTP			
CRE-CLEC-85 protein	*Caenorhabditis remanei*	41-207	0.033
YadA domain-containing protein	*Rahnella sp. Y9602*	49-192	0.075
Phage protein	*Methanosarcina vacuolata*	37-191	0.085
GG21511	*Drosophila erecta*	49-200	0.13
Glycosyl transferase family 1	*Myxococcus sp.*	72-197	0.67
PSIBLAST			
Quinolinate phosphoribosyl transferase	*Burkholderia sp. MSh2*	56-191	0.001
I-TASSER			
Myeloma immunoglobulin D delta	*Homo sapiens*		1.39
Survival motor neuron protein	*Homo sapiens*		2.59, 1.35, 2.23, 2.15
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		2.47
Myc box dependent interacting protein 1	*Homo sapiens*		1.28, 2.16
Accumulation associated protein	*Staphylococcus epidermidis*		2.24
Long tail fiber protein P37	*Enterobacteria phage T4*		1.01
Survival motor neuron protein	*Homo sapiens*		0.805
Predict Protein			
Protein binding		1, 186-187, 206	
Secreted			
DNA-directed RNA polymerase **(7)**	*Babesia bigemina*		6e-20-0.002
DNA-directed RNA polymerase **(6)**	*Phaeodactylum tricornutum*		2e-20-4e-10
DNA-directed RNA polymerase **(9)**	*Phytophthora ramorum*		4e-20-0.002
Paternally-expressed gene 3 protein **(12)**	*Bos taurus*		1e-08-5e-06
DNA-directed RNA polymerase II subunit RPB1 **(3)**	*Caenorhabditis elegans*		8e-11-3e-05
DNA-directed RNA polymerase II subunit RPB1 **(6)**	*Homo sapiens*		1e-09-3e-04

DNA-directed RNA polymerase II subunit RPB1 **(6)**	*Mus musculus*		2e-09-3e-04
DNA-directed RNA polymerase **(2)**	*Aphanomyces astaci*		1e-19, 4e-19
Atome2			
Photosystem II: Subunit PsbA	*Thermosynechococcus vulcanus*		75.97
Cytosolic leucyl-tRNA synthetase	*Candida albicans*		51.50
Cytochrome b6 **(3)**	*Mastigocladus laminosus*		37.79-24.69

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XLIII. *Lasmigona compressa* H-*ORF* sequence 2 function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		27-30	99.17
TIGR01167 LPXTG cell wall anchor domain		7-23	99.09
TIGR03304 outer membrane insertion C-terminal signal		33-34	98.82
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		112-125	97.62
TIGR03501 GlyGly-CTERM domain		7-17	97.02
CG7685-PA	*Drosophila melanogaster*	1-25	94.48
TIGR00756 pentatricopeptide repeat domain		24-44	92.77
C53B4.8	*Caenorhabditis elegans*	62-183	87.67
W02B8.6	*Caenorhabditis elegans*	62-191	83.43
W02B8.6	*Caenorhabditis elegans*	63-185	80.49
C53B4.8	*Caenorhabditis elegans*	62-185	78.70
Glycine rich protein family		4-22	78.30

Saliv_gland_allergen_Aed3		1-18	75.93
Glycine rich protein family		4-25	74.77
COLlagen family member (col-102)	*Caenorhabditis elegans*	1-45	73.01
Signal transduction histidine kinase	*Lactobacillus casei*	1-35	72.62
Secreted protein	*Streptomyces coelicolor*	5-23	72.05
LPS export ABC transporter permease LptF		1-28	71.89
Delta-notch-like EGF repeat-containing transmembrane	*Homo sapiens*	8-111	71.68
W02B8.4	*Caenorhabditis elegans*	62-185	71.01
CG13969-PA	*Drosophila melanogaster*	2-43	70.23
TonB family protein	*Nostoc punctiforme*	2-29	69.50
C46H11.8	*Caenorhabditis elegans*	5-19	68.92
CG3066-PD, isoform D	*Drosophila melanogaster*	1-26	67.62
Y81G3A.5	*Caenorhabditis elegans*	1-45	66.66
W02B8.3	*Caenorhabditis elegans*	61-191	66.37
CG18628-PA	*Drosophila melanogaster*	5-66	64.46
RCR		7-25	64.15
MoLTing defective family member (mlt-10)	*Caenorhabditis elegans*	56-185	63.65
Two component system histidine kinase	*Methanosarcina mazei*	1-39	63.59
Secreted protein	*Beggiatoa sp. PS*	1-34	63.36
Diguanylate cyclase/phosphodiesterase	*Beggiatoa sp. PS*	1-34	61.98
W02B8.3	*Caenorhabditis elegans*	61-185	61.51
R160.4	*Caenorhabditis elegans*	1-35	61.27
Rhodanese-like protein	*Beggiatoa sp. PS*	1-21	60.53
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-2)	*Caenorhabditis elegans*	5-24	60.21
Cytochrome C-type protein NapC	*Beggiatoa sp. PS*	1-34	60.08
SVM protein signal sequence		1-22	59.94
CG11020-PA, isoform A	*Drosophila melanogaster*	1-37	59.85
Synoviolin 1 isoform a	*Homo sapiens*	2-35	59.82
T27F7.3a	*Caenorhabditis elegans*	1-35	59.62
Synoviolin 1 isoform b	*Homo sapiens*	2-35	59.50
H/K_exch_ATPase_C		1-29	58.96
T19H12.3	*Caenorhabditis elegans*	5-20	58.92
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-90)	*Caenorhabditis elegans*	5-24	58.55
F59E11.5	*Caenorhabditis elegans*	1-30	58.32

Histidine kinase	*Nitrosopumilus maritimus SCM1*	1-35	58.32
TonB family protein	*Nostoc punctiforme*	1-29	58.29
DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	1-40	58.24
Temporarily Assigned Gene name family member (tag-254)	*Caenorhabditis elegans*	5-20	58.11
Synoviolin 1	*Mus musculus*	2-36	57.24
F55A11.3	*Caenorhabditis elegans*	2-29	56.91
Nuclease	*Nostoc punctiforme*	1-32	56.83
F38A3.2	*Caenorhabditis elegans*	2-45	56.31
LPS export ABC transporter permease LptG.		1-28	56.31
CG15225-PA	*Drosophila melanogaster*	6-55	56.27
Neuropeptide-Like Protein family member (nlp-16)	*Caenorhabditis elegans*	5-17	55.98
Sensory box histidine kinase PhoR	*Staphylococcus aureus*	1-35	55.95
GDSL family lipase	*Nitrosopumilus maritimus*	5-29	55.81
Predict Protein			
Protein binding		33, 38-40, 95, 97, 127, 143, 150, 173-174, 192-193	
Nucleus			
DNA-directed RNA polymerase **(5)**	*Saprolegnia parasitica*		2e-24- 5e-14
DNA-directed RNA polymerase **(5)**	*Chlamydomonas reinhardtii*		2e-24-1e-12
DNA-directed RNA polymerase **(7)**	*Phaeodactylum tricornutum*		4e-25- 2e-13
DNA-directed RNA polymerase **(5)**	*Thalassiosira oceanica*		2e-24-1e-15
DNA-directed RNA polymerase **(4)**	*Bathycoccus prasinos*		2e-24-4e-11
DNA-directed RNA polymerase II subunit RPB1 **(5)**	*Caenorhabditis briggsae*		8e-15-6e-05

DNA-directed RNA polymerase II subunit RPB1 **(3)**	*Caenorhabditis elegans*		3e-17-2e-07
DNA-directed RNA polymerase II subunit RPB1 **(6)**	*Homo sapiens*		6e-15- 9e-09
DNA-directed RNA polymerase II subunit RPB1 **(7)**	*Mus musculus*		1e-14-2e-07
Cell surface glycoprotein 1 **(10)**	*Clostridium thermocellum*		1e-04-8e-04
BLASTP/PSIBLAST			
Proteoglycan	*Histoplasma capsulatum*	49-186	5e-09
Viral protein TPX	*Histoplasma capsulatum*	53-188	1e-05
BNIP2 motif containing molecule at the carboxyl terminal region 1-like protein	*Camelus ferus*	45-186	3e-04
Rnd efflux system, outer membrane lipoprotein, NodT family	*Sodalis praecaptivus*	58-169	3e-04
Proteoglycan	*Histoplasma capsulatum*	42-186	3e-04
BLASTP			
General transcription factor 3C polypeptide 2	*Aegilops tauschii*	55-182	0.008
Histone-lysine N-methyltransferase ATXR3	*Medicago truncatula*	60-186	0.017
Histone-lysine N-methyltransferase E(z)	*Medicago truncatula*	60-186	0.018
Chitinase III	*Acanthocheilonema viteae*	45-186	0.030
Neurofilament protein	*Doryteuthis pealeii*	45-186	0.052
CRE-CLEC-85 protein	*Caenorhabditis remanei*	51-184	0.056
Adhesin	*Rahnella aquatilis*	53-186	0.25
Filamentous hemagglutinin outer membrane protein	*Stanieria cyanosphaera*	63-168	0.44
I-TASSER			
myeloma immunoglobulin D delta	*Homo sapiens*		1.36
Survival motor neuron protein **(4)**	*Homo sapiens*		2.06-1.91
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		2.52
Myc box dependent interacting protein 1 **(2)**	*Homo sapiens*		1.28, 2.25
Accumulation associated protein	*Staphylococcus epidermidis*		1.80
Restin	*Homo sapiens*		1.07
Survival motor neuron protein	*Homo sapiens*		0.748
Atome2			

Capsid protein	*Rubella virus*		83.05
Apocytochrome f	*Chlamydomonas reinhardtii*		53.92
Vpu protein	*Human immunodeficiency virus 1*		43.79
BA3-type cytochrome-c oxidase	*Thermus thermophilus*		43.07
Vpu protein **(2)**	*Human immunodeficiency virus 1*		40.31-40.28
PlnJ	*Lactobacillus plantarum*		22.71
Signal recognition 54 kDa protein	*Sulfolobus solfataricus*		22.38

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XLIV. *Lasmigona subviridis* H-*ORF* sequence 1 function predictions

Hits (n)	**Species**	**Position**	**Probability**
HHpred			
TIGR03304 outer membrane insertion C-terminal signal		1-5	99.24
TIGR04294 prepilin-type processing-associated H-X9-DG domain		34-37	99.14
TIGR01167 LPXTG cell wall anchor domain		14-30	99.03
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		95-101	97.61
TIGR03501 GlyGly-CTERM domain		14-24	96.98
CG7685-PA	*Drosophila melanogaster*	5-32	94.64
TIGR00756 pentatricopeptide repeat domain		31-51	92.73
F32A11.7	*Caenorhabditis elegans*	53-187	93.21
W02B8.6	*Caenorhabditis elegans*	60-188	93.18
MoLTing defective family member (mlt-10) **(2)**	*Caenorhabditis elegans*	65-188	92.62, 84.40
W02B8.4	*Caenorhabditis elegans*	53-188	91.89
W02B8.6	*Caenorhabditis elegans*	66-188	91.77

F32A11.7	*Caenorhabditis elegans*	66-186	91.00
W02B8.4	*Caenorhabditis elegans*	66-188	88.18
W02B8.3	*Caenorhabditis elegans*	60-188	84.25
TMEM156 protein family		7-33	83.15, 83.08
W02B8.3	*Caenorhabditis elegans*	103-188	80.80
Conserved inner membrane protein	*Escherichia coli*	1-35	80.68
Sensory box histidine kinase PhoR	*Staphylococcus aureus*	7-42	79.66
SrtB		5-40	78.93
Sensor histidine kinase	*Streptococcus pneumoniae*	7-42	78.59
LptF_YjgP LPS export ABC transporter permease LptF.		5-35	78.20
Glycine rich protein family		11-29	72.98
CG11020-PA, isoform A	*Drosophila melanogaster*	7-44	72.67
CbiN ABC-type cobalt transport system, periplasmic component		7-42	72.51
Saliv_gland_allergen_Aed3		8-25	71.19
Peptidoglycan-associated lipoprotein Pal	*Yersinia pestis*	6-25	70.56
Cytochrome c-type biogenesis protein cycj	*Bartonella henselae*	3-48	69.67
Glycine rich protein family		11-32	69.14
Urinary protein (RUP)/acrosomal protein SP-10.		3-33	68.77
DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	7-47	68.71
LPS export ABC transporter permease LptG.		5-35	67.34
TWiK family of potassium channels family member (twk-11)	*Caenorhabditis elegans*	4-29	66.04
CG13969-PA	*Drosophila melanogaster*	9-50	65.81
MORN repeat protein	*Beggiatoa sp. PS*	7-31	64.41
Signal transduction histidine kinase	*Lactobacillus casei*	8-42	63.91
Secreted protein	*Streptomyces coelicolor*	12-30	63.54
TonB family protein	*Nostoc punctiforme*	9-36	63.16
C53B4.8	*Caenorhabditis elegans*	65-185	62.28
H/K_exch_ATPase_C		7-36	62.23
Lipoprotein required for capsular polysaccharide translocation through the outer membrane	*Escherichia coli*	7-25	61.73
R160.4	*Caenorhabditis elegans*	7-42	61.58
SVM protein signal sequence		8-29	61.27
Nitric oxide reductase subunit C; metal-binding, membrane protein, immune system-oxidoreductas	*Pseudomonas aeruginosa*	5-32	60.51

COLlagen family member (col-102)	*Caenorhabditis elegans*	7-52	60.47
Cytoplasmic membrane protein	*Bartonella henselae*	7-40	60.06
CG3066-PD, isoform D	*Drosophila melanogaster*	8-33	59.22
Synoviolin 1 isoform a	*Homo sapiens*	7-42	59.11
Protein-export membrane protein	*Agrobacterium tumefaciens*	6-30	58.68
CG7875-PA	*Drosophila melanogaster*	2-44	58.12
Synoviolin 1 isoform b	*Homo sapiens*	7-42	57.82
C46H11.8	*Caenorhabditis elegans*	12-26	57.66
Permease YjgP/YjgQ family protein	*Nostoc punctiforme*	5-36	57.45
RCR		14-32	57.43
F55A11.3	*Caenorhabditis elegans*	7-36	57.42
BLASTP/PSIBLAST			
Herpes virus major outer envelope glycoprotein (BLLF1)	*Herpes virus*	69-195	2.73e-03
small proline-rich protein 3	*Mus musculus*	69-195	3e-04
ARF GAP-like zinc finger-containing protein	*Trichomonas vaginalis*	45-194	4e-04
aggrecan core protein precursor	*Sus scrofa*	57-192	8e-04
viral protein TPX	*Histoplasma capsulatum*	64-192	0.003
BLASTP			
Proteoglycan **(3)**	*Histoplasma capsulatum*	59-192	0.008-0.048
Small proline-rich protein 3 **(2)**	*Rattus norvegicus*	68-197	0.033
FecR protein	*Cylindrospermum stagnale*	53-195	0.065
Peptidase S8	*Actinobacillus capsulatus*	71-193	0.10
Aggrecan core protein	*Bos mutus*	62-192	0.56
CD5 antigen-like protein	*Chelonia mydas*	54-191	1.00
I-TASSER			
Chaperone protein PAPD	*Escherichia coli*		1.13
Survival motor neuron protein **(2)**	*Homo sapiens*		1.89, 1.92
Type I hyperactive antifreeze protein **(2)**	*Pseudopleuronectes americanus*		1.77, 1.18
Myc box dependent interacting protein 1 **(2)**	*Homo sapiens*		1.04, 2.07
Major capsid protein	*Synechococcus phage Syn5*		1.66
Type I hyperactive antifreeze protein	*Pseudopleuronectes americanus*		0.767
Dynamin family protein	*Nostoc punctiforme*		0.600

Phospholipase C beta	*Meleagris gallopavo*		0.574
LEOA	*Escherichia coli*		0.573
Interferon-induced guanylate-binding protein 1	*Homo sapiens*		0.571
TcdA1	*Photorhabdus luminescens*		0.563
Tyrosine-protein kinase Fes/Fps	*Homo sapiens*		0.559
RhUL123	*Macacine herpesvirus 3*		0.547
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-3	*Homo sapiens*		0.545
Predict Protein			
Protein binding		37, 50-51, 63, 124, 151-153, 169-171	
Cytoplasm			
Merozoite surface protein 1	*Plasmodium reichenowi*		4e-18
Protein piccolo **(5)**	*Rattus norvegicus*		4e-04-0.35
Merzoite surface protein 1	*Plasmodium reichenowi*		6e-22, 1e-08
Cell surface glycoprotein 1 **(21)**	*Clostridium thermocellum*		0.011-0.85
Atome2			
Protein kinase BYR2	*Schizosaccharomyces pombe*		51.37
Cell wall surface anchor family protein	*Streptococcus pneumoniae*		46.70
Apocytochrome f	*Chlamydomonas reinhardtii*		33.57
Cytochrome B6 **(3)**	*Mastigocladus laminosus*		23.23-19.48
Bone marrow stromal antigen 2	*Homo sapiens*		22.50
30S ribosomal protein S27E	*Archaeoglobus fulgidu*		21.44
SERCA1a	*Oryctolagus cuniculus*		20.68

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of

10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.

Supplementary Table XLV. *Lasmigona subviridis* H-*ORF* sequence 2 function predictions

Hits (n)	Species	Position	Probability
HHpred			
TIGR04294 prepilin-type processing-associated H-X9-DG domain		28-31	99.07
TIGR01167 LPXTG cell wall anchor domain		8-24	98.98
TIGR03304 outer membrane insertion C-terminal signal		91-93	98.86
TIGR03057 X-X-X-Leu-X-X-Gly heptad repeats		91-95	97.52
TIGR03501 GlyGly-CTERM domain		8-18	96.88
CG7685-PA	*Drosophila melanogaster*	1-26	94.13
W02B8.6	*Caenorhabditis elegans*	54-220	93.06
TIGR00756 pentatricopeptide repeat domain		25-45	92.04
MoLTing defective family member (mlt-10)	*Caenorhabditis elegans*	59-219	90.37
W02B8.6	*Caenorhabditis elegans*	60-219	89.42
F32A11.7	*Caenorhabditis elegans*	47-219	87.42
W02B8.4	*Caenorhabditis elegans*	47-220	87.08
F32A11.7	*Caenorhabditis elegans*	98-219	85.50
W02B8.4	*Caenorhabditis elegans*	60-218	85.32
W02B8.3	*Caenorhabditis elegans*	54-226	81.65
Sensory box histidine kinase PhoR	*Staphylococcus aureus*	1-36	77.99
Sensor histidine kinase	*Streptococcus pneumoniae*	1-36	75.75
Glycine rich protein family		5-26	73.02
MoLTing defective family member (mlt-10)	*Caenorhabditis elegans*	93-227	72.96
LPS export ABC transporter permease LptF.		1-29	72.07
Saliv_gland_allergen_Aed3		2-19	69.33
Glycine rich protein family		5-26	68.78
CbiN ABC-type cobalt transport system, periplasmic component		1-36	68.50
Conserved inner membrane protein	*Escherichia coli*	1-29	67.92
CG11020-PA, isoform A	*Drosophila melanogaster*	1-38	66.15

DumPY: shorter than wild-type family member (dpy-5)	*Caenorhabditis elegans*	1-41	64.12
CG13969-PA	*Drosophila melanogaster*	3-44	63.05
Secreted protein	*Streptomyces coelicolor*	6-24	62.50
Bromodomain adjacent to zinc finger domain, 2A	*Homo sapiens*	47-220	61.40
TonB family protein	*Nostoc punctiforme*	3-30	61.29
SVM protein signal sequence		2-23	61.17
Lipoprotein required for capsular polysaccharide translocation through the outer membrane	*Escherichia coli*	1-19	60.26
Signal transduction histidine kinase	*Lactobacillus casei*	2-36	60.14
MORN repeat protein	*Beggiatoa sp. PS*	1-25	59.96
F55A11.3	*Caenorhabditis elegans*	1-30	59.87
CG3066-PD, isoform D	*Drosophila melanogaster*	2-27	59.58
Synoviolin 1 isoform a	*Homo sapiens*	1-36	59.21
LPS export ABC transporter permease LptG.		1-29	58.32
Synoviolin 1 isoform b	*Homo sapiens*	1-36	58.28
RCR		8-26	58.05
W02B8.3	*Caenorhabditis elegans*	135-220	57.96
Y81G3A.5	*Caenorhabditis elegans*	2-46	57.87
Urinary protein (RUP)/acrosomal protein SP-10		1-27	57.72
COLlagen family member (col-102)	*Caenorhabditis elegans*	2-46	57.55
Diguanylate cyclase/phosphodiesterase	*Beggiatoa sp. PS*	2-35	56.03
T27F7.3a	*Caenorhabditis elegans*	2-36	55.88
C46H11.8	*Caenorhabditis elegans*	6-20	55.51
Synoviolin 1	*Mus musculus*	1-37	54.77
C53B4.8	*Caenorhabditis elegans*	97-217	54.61
Ttwo component system histidine kinase	*Methanosarcina mazei*	2-40	53.60
H/K_exch_ATPase_C		1-30	53.59
CG15225-PA	*Drosophila melanogaster*	7-56	52.77
Rhodanese-like protein	*Beggiatoa sp. PS*	2-22	52.41
Cytochrome C-type protein NapC	*Beggiatoa sp. PS*	2-35	51.59
Cytochrome c family protein	*Beggiatoa sp. PS*	1-23	51.54
Cytoplasmic membrane protein	*Bartonella henselae*	1-34	51.46
Prion-like-(Q/N-rich)-domain-bearing protein family member (pqn-2)	*Caenorhabditis elegans*	6-25	51.39
UCP031802		1-40	51.24
Secreted protein	*Beggiatoa sp. PS*	1-30	51.14

R160.4	*Caenorhabditis elegans*	1-36	51.13
BLASTP			
Peptidase S8	*Actinobacillus capsulatus*	70-230	0.078
Aggrecan	*Bos mutus*	85-224	0.081
Aggrecan **(5)**	*Bos taurus*	56-224	0.046-0.27
CD5 antigen-like protein	*Chelonia mydas*	48-223	0.38
Protein FAM186A, partial	*Picoides pubescens*	48-227	0.57
Small proline-rich protein 3	*Mus musculus*	158-225	1.0
BLASTP/PSIBLAST			
Herpes virus major outer envelope glycoprotein (BLLF1)	*Herpes virus*	40-227	4.85e-04
aggrecan core protein precursor	*Sus scrofa*	51-224	5e-08
RTX toxin RtxA	*Vibrio cholerae*	62-172	3e-04
proteoglycan	*Histoplasma capsulatum*	53-223	3e-04
RTX toxin	*Vibrio cholerae*	62-172	3e-04
RTX toxin RtxA	*Vibrio cholerae*	62-204	3e-04
PSIBLAST			
RTX toxin RtxA	*Vibrio cholerae*	62-172	4e-04
peptidase C80 **(3)**	*Vibrio cholerae*	62-172	4e-04-5e-04
RTX toxin RtxA	*Vibrio cholerae*	49-172	6e-04
RTX toxin, partial	*Vibrio cholerae*	62-172	8e-04
proteoglycan	*Histoplasma capsulatum*	53-223	8e-04
viral protein TPX	*Histoplasma capsulatum*	58-224	0.001
peptidase C80	*Vibrio cholerae*	62-204	0.002
ACAN protein	*Homo sapiens*	63-223	0.003
Aggrecan core protein isoform 1 precursor	*Homo sapiens*	63-223	0.003
Aggrecan core protein isoform 2 precursor	*Homo sapiens*	63-223	0.003, 0.004
FecR protein	*Cylindrospermum stagnale*	47-230	0.003
Large aggregating cartilage proteoglycan core protein	*Homo sapiens*	63-223	0.004
PSIBLAST			
RTX toxin **(9)**	*Vibrio vulnificus*	62-172	3e-04-0.001
Peptidase C80 **(5)**	*Vibrio vulnificus*	62-172	0.001
RTX toxin, partial	*Vibrio vulnificus*	62-172	0.001
RTX toxin RtxA **(6)**	*Vibrio vulnificus*	62-172	0.001-

			0.002
Peptidase C80 **(19)**	*Vibrio cholerae*	62-172	0.001-0.004
RTX toxin RtxA domain protein	*Vibrio cholerae*	62-172	0.001
RTX toxin, partial	*Vibrio vulnificus*	62-172	0.001
RTX toxin RtxA **(5)**	*Vibrio cholerae*	62-172	0.002-0.003
Autotransporter adhesin	*Vibrio ordalii*	62-172	0.002
PGAP1-like family protein	*Vibrio cholerae*	62-172	0.002
ARF GAP-like zinc finger-containing protein	*Trichomonas vaginalis*	72-226	0.003
RTX toxins determinant A	*Vibrio cholerae*	62-172	0.003
rtxA repeat family protein **(2)**	*Vibrio cholerae*	49-172	0.003
RTX toxins determinant A and related Ca2+-binding proteins	*Vibrio cholerae*	62-172	0.003
Peptidase C80	*Photorhabdus asymbiotica*	62-172	0.003
Peptidase C80	*Photorhabdus luminescens*	62-172	0.003
RTX (Repeat in toxin) cytotoxin	*Vibrio albensis*	49-172	0.004
I-TASSER			
DNA (cytosine-5)-methyltransferase 1	*Zea mays*		1.21, 1.20
Survival motor neuron protein **(3)**	*Homo sapiens*		1.15-2.72
Short tail fiber protein	*Enterobacteria phage T4*		2.14
Myc box dependent interacting protein 1 **(2)**	*Homo sapiens*		1.09, 2.12
Major capsid protein	*Synechococcus phage Syn5*		2.19
Accumulation associated protein	*Staphylococcus epidermidis*		1.80
Major capsid protein	*Synechococcus phage Syn5*		0.714
T7-like capsid protein	*Prochlorococcus phage P-SSP7*		0.559
Coat protein	*Enterobacteria phage P22*		0.520
Phage-related protein	*Bordetella bronchiseptica*		0.511
Predict Protein			
Protein binding		1, 229	
Cytoplasm			

DNA polymerase **(6)**	*Macaca fascicularis*		7e-26-4e-22
Proteoglycan 4 **(7)**	*Mus musculus*		1e-04-0.15
Cell surface glycoprotein 1 **(20)**	*Clostridium thermocellum*		0.18-0.12
SH3 domain-containing protein C23A1.17 **(4)**	*Schizosaccharomyces pombe*		5e-05- 0.28
Atome2			
30S ribosomal protein S27E	*Archaeoglobus fulgidus*		75.63
Cytochrome b6 **(5)**	*Mastigocladus laminosus*		72.09-52.50
Bone marrow stromal antigen 2	*Homo sapiens*		48.08
SERCA1a	*Oryctolagus cuniculus*		41.59
Serine/threonine-protein kinase MARK2	*Homo sapiens*		23.04

NOTE : For Supplementary Tables XI-XLV HHpred and @tome 2 use probabilities /100; I-TASSER Z scores > 1.0 and TM scores > 0.5 are significant; BLASTP & Predict Protein E-values < 1.0 are significant (adjusted from developer's recommendation of 10.0); PSIBLAST E-values < 0.001 are significant; Motif Scan N-scores do not have a significance threshold, a higher score indicates a better match.