Horizontally transferred genes in the ctenophore *Mnemiopsis leidyi* encode enzymes and are expressed during early development (#22476)

First submission

Editor guidance

Please submit by **29 Dec 2017** for the benefit of the authors (and your $200 publishing discount).

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data. Download from the location described by the author.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files
Download and review all files from the materials page.

4 Figure file(s)
2 Table file(s)
1 Other file(s)

Structure and Criteria

2

Structure your review

The review form is divided into 5 sections.
Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- Data is robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.
- Speculation is welcome, but should be identified as such.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Horizontally transferred genes in the ctenophore *Mnemiopsis leidyi* encode enzymes and are expressed during early development

Alexandra M Hernandez [1, 2] , **Joseph F Ryan** [Corresp. 1, 2]

[1] Whitney Laboratory for Marine Bioscience, St. Augustine, Florida, United States
[2] Department of Biology, University of Florida, Gainesville, FL, United States

Corresponding Author: Joseph F Ryan
Email address: joseph.ryan@whitney.ufl.edu

Horizontal gene transfer has had major impacts on the biology of a wide range of organisms from antibiotic resistance in bacteria to adaptations to herbivory in arthropods. A growing body of literature shows that horizontal gene transfer (HGT) between non-animals and animals is more commonplace than previously thought. In this study, we present a thorough investigation of HGT in the ctenophore *Mnemiopsis leidyi*. We applied tests of phylogenetic incongruence to identify nine genes that were likely transferred horizontally early in ctenophore evolution from bacteria and non-metazoan eukaryotes. All but one of these HGTs (an uncharacterized protein) appear to perform enzymatic activities in *M. leidyi*, supporting previous observations that enzymes are more likely to be retained after HGT events. We found that the majority of these nine horizontally transferred genes were expressed during early development, suggesting that they are active and play a role in the biology of *M. leidyi*. This is the first report of HGT in ctenophores, and contributes to an ever-growing literature on the prevalence of genetic information flowing between non-animals and animals.

Horizontally transferred genes in the ctenophore *Mnemiopsis leidyi* encode enzymes and are expressed during early development

Alexandra M. Hernandez[1,2] and Joseph F. Ryan[1,2]

[1] Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL, USA

[2] Department of Biology, University of Florida, Gainesville, FL, USA

Corresponding Author:

Joseph Ryan[1]

joseph.ryan@whitney.ufl.edu

Abstract

Horizontal gene transfer has had major impacts on the biology of a wide range of organisms from antibiotic resistance in bacteria to adaptations to herbivory in arthropods. A growing body of literature shows that horizontal gene transfer (HGT) between non-animals and animals is more commonplace than previously thought. In this study, we present a thorough investigation of HGT in the ctenophore *Mnemiopsis leidyi*. We applied tests of phylogenetic incongruence to identify nine genes that were likely transferred horizontally early in ctenophore evolution from bacteria and non-metazoan eukaryotes. All but one of these HGTs (an uncharacterized protein) appear to perform enzymatic activities in *M. leidyi,* supporting previous observations that enzymes are more likely to be retained after HGT events. We found that the majority of these nine horizontally transferred genes were expressed during early development, suggesting that they are active and play a role in the biology of *M. leidyi*. This is the first report of HGT in ctenophores, and contributes to an ever-growing literature on the prevalence of genetic information flowing between non-animals and animals.

Introduction

Evolution is commonly thought to occur by descent with modification from a single lineage. However, evidence has shown that genomes from bacteria, archaea, and eukaryotes are typically chimeric, resulting from horizontal (or lateral) gene transfers (Garcia-Vallvé et al. 2000; Katz 2002). As such, horizontal gene transfer (HGT) has likely impacted evolution more than originally thought by creating opportunities for rapid genetic diversification and contributing to speciation events. Moreover, HGT is a potential catalyst for organisms to acquire novel traits (Soucy et al. 2015) and creates opportunities for HGT receivers to exploit new ecological niches (Boto 2010). For example, HGTs have played an important role in herbivory in arthropods (Wybouw et al. 2016), venom recruitment in parasitoid wasps (Martinson et al. 2016), cellulose production in urochordates (Dehal et al. 2002) and plant parasitism in nematodes (Haegeman et al. 2011).

Although HGT is generally accepted as an important evolutionary mechanism in prokaryotes (Boto 2014), it remains controversial whether it occurs in animals, despite many convincing studies (Madhusoodanan 2015). Much of the skepticism has been fueled by high-profile reports of HGT (e.g. Lander et al. 2001; Boothby et al. 2015) that were later shown to be largely incorrect (Stanhope et al. 2001; Koutsovoulos et al. 2016). In addition, HGT in animals is hypothesized to be rare due to the origin of a sequestered germ line, which provides less opportunities for germ cells to be exposed to foreign DNA (Doolittle 1999; Andersson et al. 2001; Jensen et al. 2016). However, the presence and absence of germline sequestration is not well described across the animal tree of life, and there are inconsistencies between studies regarding which animal groups have sequestered germlines (Buss, 1983; Radzvilavicius et al. 2015; Jensen et al. 2016).

46 The major challenges for HGT detection efforts have been taxon sampling and
47 contamination. Many early reports of HGT in animals were overturned due to limited
48 representation of taxa in public genomic databases (e.g. Salzberg et al. 2001). For example, a
49 gene present in bacteria and humans, but absent from nematodes and drosophilids (the most
50 highly represented taxa at the time) may have been considered the result of HGT, until
51 discovering that the gene is present in many other animal genomes that were not available at the
52 time of the initial claim. In these cases, the limited representation of taxa made it difficult to
53 distinguish HGTs from differential gene loss (Andersson 2006 et al. 2006; Keeling & Palmer
54 2008). More recently, contamination has led to both overestimation and likely underestimation of
55 HGT events. In several recent cases, contamination in newly generated datasets has been
56 interpreted as HGT but later shown to be cross-contaminants present in genome sequences
57 (Bhattacharya et al. 2013; Delmont & Eren 2016; Koutsovoulos et al. 2016). On the other hand,
58 the presence of contaminants in public databases (e.g., a bacteria sequence labeled as an animal
59 sequence) makes it difficult to identify *bona fide* HGTs, as "animal" sequences will appear
60 among the top BLAST hits for a particular HGT, leading to false negatives. As such,
61 contamination remains a major hurdle to contemporary studies of HGT.

62 Pairwise BLAST-based similarity scores (e.g. alien index (Gladyshev et al. 2008) and the
63 HGT index (Boschetti et al. 2012)) are the most common criteria used to detect HGT in animals.
64 However, these measures largely ignore phylogenetic information associated with sequence data.
65 While a positive BLAST-based result may be due to HGT, it may also result from gene loss,
66 selective evolutionary rates, convergent evolution, sequence contamination, and species
67 misassignment (Hall et al. 2005). Previous HGT studies have demonstrated that HGT predictions
68 need to be carefully considered and a combination of methods are required to rule out false

positives (Schönknecht et al. 2013). Hypothesis tests incorporating phylogenetic incongruence are one such method that has been used to test HGT. While some studies in animals have incorporated these techniques (e.g. Eliáš et al, 2016), they are more commonly deployed in studies involving non-animals (e.g. Bapteste et al. 2003; Richards et al. 2006).

HGT has yet to be thoroughly explored in Ctenophora. Ctenophores (comb jellies) are marine invertebrates that are morphologically characterized by eight rows of cilia used for movement. They typically live in the water column, but the group includes benthic species as well (Song & Hwang 2010; Alamaru et al. 2015; Glynn et al. 2017). Evidence has suggested that ctenophores are the sister group to all other animals (Dunn et al. 2008; Hejnol et al. 2009; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Torruella et al. 2015; Whelan et al. 2015; Arcila et al. 2017; Shen et al. 2017; Whelan et al. 2017), but the position remains controversial (Pisani et al. 2015, Simion et al. 2017). Thus, investigating HGT in ctenophores is essential to understanding its implications on early animal evolution.

Here, we apply a rigorous framework to identify and confirm likely HGTs in the ctenophore *Mnemiopsis leidyi.* Our process includes identification of HGT candidates by alien index and confirmation by phylogenetic hypothesis testing, providing statistical support in an evolutionary framework. Furthermore, we analyze gene expression profiles during early development to obtain clues as to the function of these HGTs in *M. leidyi*.

Material and Methods

All command lines, parameters, and version numbers of programs are in the supplementary text.

Identification of candidate HGTs by alien_index

As part of this project, we developed the program alien_index and complimentary metazoan/non-metazoan sequence databases to automate the generation of alien index (Gladyshev et al. 2008) and HGT index scores (Boschetti et al. 2012). We BLASTed a database of animal and non-animal sequences (alien_index_db version 0.01) and then calculated alien index values as the logarithmic difference between the best BLASTP E-values for animal and non-animal hits (as outlined in Gladyshev et al. 2008). This database includes translated gene models from curated genomes that include bacteria (5), archaea (2), non-animal eukaryotes (5), and animals (11). See Table S1 or http://ryanlab.whitney.ufl.edu/downloads/alien_index/ for the entire list of taxa. HGT index values were computed by the difference in the highest non-metazoan and metazoan bit scores generated from the alien_index database. The alien_index program is available at: https://github.com/josephryan/alien_index

Confirmation of likely HGTs

We verified that HGT candidates identified by alien_index were not the result of bacterial contaminants by using the *M. leidyi* genome browser (Moreland et al. 2014) to examine the intron/exon structure of each HGT candidate, as well as the origin of their neighboring genes. We also searched for each HGT candidate (identified from the genome and gene models from an *M. leidyi* individual collected in Woodshole, MA) in the transcriptome of an *M. leidyi* individual collected from St. Augustine, Florida, as well as in seven other ctenophore transcriptomes reported in Moroz et al. (2014): *Bolinopsis infundibulum, Beroe abyssicola, Dryodora glandiformis, Pleurobrachia bachei, Vallicula multiformis, Coeloplana astericola, Euplokamis dunlapae*.

Once HGT candidates were filtered for contaminants, we performed maximum-likelihood analyses on putative HGTs to confirm non-animal origin. HGT candidates were used as queries for BLASTP against NCBI's RefSeq database (O'Leary et al. 2016) using the NCBI BLAST interface. We collected the top ten sequences each from bacteria, eukaryotes, fungi, and animals with an E-value cutoff of 0.1. We included only the first sequence if there were hits to sequences from species in the same genus. We also added sequences from *Amphimedon queenslandica, Trichoplax adhaerens*, *Nematostella vectensis*, *Capitella teleta*, *Drosophila melanogaster*, and *Homo sapiens* from version 0.01 of the alien_index database that fit the above criteria. Sequences were aligned against the corresponding putative HGT using MAFFT (Katoh et al. 2002; Katoh & Standley 2013) and trimmed with Gblockswrapper (Castresana 2000). There were six genes without animal BLASTP hits (E-value ≤ 0.1), which prevented us from performing additional phylogenetic analyses. We considered the lack of animal BLASTP hits below our cutoff as sufficient evidence that these six were clearly HGTs. ML018031a only had two BLASTP hits to animal sequences. Since it was unclear if this resulted from contamination, we were unable to test this gene using phylogenetic approaches, so it was removed from contention as an HGT.

We performed maximum-likelihood analyses for each alignment using RAxML version 8.1.21 (Stamatakis 2014). Since the RefSeq database has many instances of contamination (Pible et al. 2014), we allowed a maximum of two non-ctenophore animal sequences to fall outside of the main animal clade. To implement this, we pruned putative contaminants if the removal of two taxa resulted in a monophyletic animal clade (Fig. S1). We discarded any HGT candidates with more than two taxa disrupting animal monophyly.

We explicitly tested topologies in opposition to HGT (i.e. animal monophyly) with the SOWH test using SOWHAT (Church et al. 2015) and the AU test using CONSEL (Shimodaira and Hasegawa 2001). The SOWH and AU test evaluate statistical support for phylogenetic incongruence by comparing the likelihood values between trees to a distribution of trees generated by parametric sampling in the SOWH test and non-parametric sampling in the AU test. To address any potential problems of selection bias in the AU test (causing the likelihood value to bias upwards for the maximum likelihood best tree when included in the dataset), we performed multiple AU analyses using bootstrap trees as suboptimal trees (similar to Eliáš et al. 2016). We generated 100 bootstrap trees using RAxML rapid bootstrap analyses, and verified there were no duplicate trees in our 100 bootstrap set using the ape package in R (Paradis et al. 2004). RAxML was used to generate per-site log likelihoods for the best maximum likelihood tree, the tree constraining the putative HGT to metazoans, and suboptimal trees, for input in CONSEL. To test the effectiveness of comparing to bootstrap trees, we manually created a set of suboptimal trees for each candidate HGT by shuffling clades of three (Fig. S2) and running the same analyses. We evaluated the tree space covered by suboptimal trees in the AU test (i.e. bootstrap and manually generated trees) by visualizing the data using violin plots. We calculated likelihood proportions for each tree by dividing individual likelihood scores by the average likelihood score of suboptimal trees. The trees and scripts used to automate these phylogenetic analyses are available in the accompanying GitHub site.

HGT developmental expression profiles

An extensive early developmental transcriptome of *Mnemiopsis leidyi* was recently generated from single-embryo RNA-Seq analyses for developmental stages during the first 20 hours (Levin et al. 2016). To these expression profiles we added six additional time points (four

replicates each) for hours 14-19. These additional data were produced by Itai Yanai and Mark Martindale using the methods outlined in Levin et al. (2016). We summed median transcript-per-million values along the 25 time points for each of our 9 confirmed HGTs. HGTs that had summed median read counts of 100 or greater were classified as being expressed sufficiently to have roles in early development.

HGT origins and functions

We determined the origin of likely HGTs by using the NCBI BLAST interface for BLASTP of HGTs against NCBI's RefSeq database. Metazoans were excluded from these BLASTP searches and we recorded the origin of the top hits (E-value ≤0.1). To uncover the functional roles of HGTs, we used the BLAST interface provided by UniProt and the UniProtKB database (Pundir et al. 2017). Annotations of the top hits (E-value ≤0.1) were transferred to HGT candidates. We also associated HGTs with Pfam-A domains using the MGP Portal under the *Mnemiopsis* Gene Wiki (Moreland et al. 2014).

Results

Mnemiopsis leidyi HGTs

We calculated an alien index for every *M. leidyi* gene model using a database of 11 animals and 12 non-animals (Table S1). We identified 37 genes with alien indices greater than 45 and designated these as HGT candidates (cut-off values were established by Gladyshev et al. (2008)). We used the *M. leidyi* genome browser to examine the intron/exon structure of each HGT candidate, as well as the origin of their neighboring genes for evidence of bacterial

contamination (lack of introns would indicate bacterial contamination). All but one HGT candidate, were found on scaffolds with intron-containing genes and 73% of the candidates had introns (Table S2). The only exception, ML49231a (itself containing 6 introns) was the only gene on its scaffold. These data suggest that our HGT candidates are not bacterial contaminants.

We confirmed each of the HGT candidates in a transcriptome from an *M. leidyi* individual collected in St. Augustine, FL (*M. leidyi* genome and gene models were from individuals collected in Woods Hole, MA). We also searched for each HGT candidate in seven ctenophore transcriptomes published in Moroz et al. (2014): *Bolinopsis infundibulum, Beroe abyssicola, Dryodora glandiformis, Pleurobrachia bachei, Vallicula multiformis, Coeloplana astericola, Euplokamis dunlapae*. Each HGT candidate was present in the transcriptome of at least one other ctenophore species and in the Florida *M. leidyi* transcriptome (Fig. 1). Because it is unlikely that the same species contaminated each of these datasets, these comparisons provide additional evidence against these sequences being the result of contamination.

In addition to the alien_index database, we BLASTed the RefSeq database at NCBI restricting hits to bacteria, then to animals, and then to non-animal eukaryotes. All but six HGT candidates had BLAST hits to animals with E-values ≤ 0.1. We counted these six (ML012034a, ML06718a, ML03277a, ML02232a, ML18354a, ML219316a) as likely HGTs and performed additional investigations of the remaining 30 HGT candidates.

For the remaining 30 candidates, we conducted detailed phylogenetic analyses using the top 10 hits of unique non-animal and animal taxa from each of the RefSeq searches along with sequences from *Amphimedon queenslandica, Trichoplax adhaerens*, *Nematostella vectensis*, *Capitella teleta*, *Drosophila melanogaster*, and *Homo sapiens* that were top hits from our initial

BLASTs of the alien_index database. Candidate HGTs that formed a clade with all other animals were ruled out as potential HGTs and candidates that disrupted animal monophyly were tested further. We discarded candidates with more than two non-ctenophore animal sequences disrupting animal monophyly; in the case of two or less sequences the disrupting sequences were considered potential contaminants and pruned (e.g. Fig. S1). We then applied the SOWH and AU tests to the remaining candidates to compare the maximum likelihood topology to the alternative hypothesis that candidate HGTs were more closely related to animals (Fig. 2). This involved comparing likelihood values of optimal trees to those that were constrained to produce a monophyletic Animalia. Our results showed that the AU test was more conservative in confirming HGTs than the SOWH test (Table 1). For perspective on how optimal trees compared to constrained trees, we ran AU tests comparing optimal trees to bootstrap trees (sub-optimal trees covering a wide range of tree space) (Fig. 3). The likelihood scores of the constrained trees from our confirmed HGTs in the AU test tend to fall outside or on the tails of the distribution of likelihood scores of suboptimal trees, whereas the likelihood scores of constrained trees for unconfirmed HGTs were all closer to the most likely tree than the bootstrap trees (Fig. 3). We confirmed seven HGTs in which gene trees significantly differed ($p < 0.05$) from the metazoan constraint trees in both the SOWH and AU analyses (Table 1). This brought our total to 13 likely HGTs.

Although 13 HGTs were verified by BLAST and phylogenetic analyses, we removed 4 of these from contention (ML092610a, ML06718a, ML03277a, ML02232a) because the top hits from BLAST against RefSeq were either Choanoflegellida or Ichythosporea (groups closely related to animals). If ctenophores are the sister group to the rest of animals, vertical inheritance remains a possibility for these cases. As such, we confirm a total of nine highly likely HGTs.

HGTs are expressed in early development

We summed transcript-per-million values (medians for each set of expression values at 25 time points) from single-embryo RNA-Seq analyses over 20 hours for each of the nine confirmed HGTs. Six of the nine HGTs had sums greater than 100 (Fig. 4), suggesting that these had some role in early development.

HGTs are enzymes originating from non-animal eukaryotes and bacteria

We determined the origin of the nine confirmed highly likely HGTs by using BLAST against the RefSeq database. Four HGTs appear to have originated from bacteria and five from non-animal eukaryotes (Table 2). The five HGT events that originated from non-animal eukaryotes appear to be from Stramenopiles, Viriplantae, Rhodophyta or Amoebazoa. To characterize gene function, we BLASTed the nine confirmed HGTs against the UniProt database. All HGTs except one uncharacterized protein (ML219316a) appear to be an enzyme and/or have domains that perform catalytic functions (Table 2).

Discussion

HGTs in ctenophores and their implications

It had been speculated previously that ctenophores had HGTs since initial profiling revealed that many ‘bacteria-like’ genes in ctenophores contain introns and should be on chromosomes with vertically inherited (i.e. non-HGT) genes (Artamonova et al. 2015). We identified 37 HGT candidates by using alien_index. Evidence from our study confirmed that 73% of HGT candidates had introns and all but one gene (the only gene on this scaffold) were on

scaffolds with other intron-containing genes (Table S2). This provided evidence that these candidates were unlikely the result of extrinsic contamination. We provided additional evidence that candidates did not result from contamination by showing that all HGTs were found in both Massachusetts and Florida *M. leidyi* individuals as well as many other ctenophore species (Fig. 1). Six HGTs are present in the *E. dunlapae* transcriptome suggesting that the majority of these HGT events occurred very early in ctenophore evolution (Fig. 1). This deep evolutionary history suggests that these HGTs may have had important impacts on the biology of ctenophores.

Mechanisms driving HGT in ctenophores

While we are uncertain about the mechanisms driving HGT, we speculate that some of these may have resulted from symbiotic relationships with bacteria and non-animal eukaryotes. *Gammaproteobacteria* and *Bacteroidetes* have been identified as two of the most abundant groups of bacteria associated with ctenophores (Daniels & Breitbart 2012). These groups were identified as the likely donors of three HGTs (i.e. ML00955a, ML02771a, ML18354a) in the *M. leidyi* genome and confirmed in almost all other ctenophores transcriptomes. Other possible donors could have been gymnamoebae symbionts that have been described living on the surface of comb plates and on the ectoderm of ctenophores (Moss et al. 2001). However, studies investigating symbiotic relationships with ctenophores are limited. Further studies are needed to improve our understanding of the impacts of symbiotic relationships on HGT, as well as to potentially understand the mechanisms that drive HGT between organisms.

***Mnemiopsis leidyi* HGTs are expressed during early development and are disproportionately enzymes**

Many HGTs are likely to be deleterious and lost, but some HGTs will be neutral or provide a selective advantage and spread throughout a population (Thomas & Nielsen 2005). HGT integration is thought to mainly occur in neutral genes with low levels of expression (Park & Zhang 2012). Once integrated, neutral HGTs may become a source of novel genetic variation upon which selection can act (Soucy et al. 2015). HGTs may then become more highly expressed after recruitment of transcription factors and regulators from the host genome (Lercher & Pál 2008). Six of the nine HGTs we identified showed high expression during the first 20 hours of development, suggesting potentially important developmental roles. ML02771a is highly expressed during early development and encodes penicillin acylase or amidase, which catalyzes the hydrolysis of benzylpenicillin. This reaction creates key intermediates for penicillin synthesis and may be important to defend against microbial infection or colonization.

Observations of HGT patterns in prokaryotes have also suggested that there is a preference to retain operational (metabolic) genes rather than informational genes (Lawrence & Roth 1996; Jain et al. 1999; Garcia-Vallvé et al. 2000). Genes involved in DNA replication, transcription, and translation are infrequently identified in sets of HGTs (Thomas & Nielsen 2005). Preference for operational genes is hypothesized to occur because informational genes are involved in larger and complex systems (Jain et al. 1999). Recently, this pattern has also been observed in animal HGTs (Boto, 2014) (e.g. Zhu et al. 2011; Boschetti et al. 2012; Sun et al. 2012; Eyres et al. 2015; Conaco et al. 2016). These reports suggest that operational genes are preferentially transferred and/or retained in both prokaryotes and eukaryotes. Our data support this idea since all of the characterizable genes in our HGT set are enzymes.

Commonly used BLAST-based methods for identifying HGTs in animals are insufficient

283 Identifying HGTs can be challenging due to bacterial associations with hosts
284 (Artamonova & Mushegian 2013; Chapman et al. 2010; Fraune & Bosch 2007), DNA extraction
285 kits and reagents that have led to contamination (Naccache et al. 2013; Salter et. al 2014), and/or
286 laboratory conditions during DNA extraction (Laurence et al. 2014; Strong et al. 2014). These
287 challenges associated with sequencing and assembly have led to contamination in public
288 databases (Longo et al. 2011; Merchant et al. 2014) and make HGT predictions difficult.
289 Moreover, while BLAST-based approaches (i.e., alien index and the HGT index) are useful for
290 identification of HGT candidates, they are difficult to implement, lack evolutionary perspective,
291 and do not address problems associated with contamination.

292 To overcome some of these challenges, we developed alien_index to automate the
293 generation of alien index and HGT index scores for rapid identification of HGT candidates. We
294 confirmed HGTs by using rigorous phylogenetic approaches to address the problems associated
295 with the lack of evolutionary perspective from BLAST methods. Our phylogenetic tests of
296 incongruence provided clear metrics from which to judge the level of certainty applied to each
297 HGT candidate. Our study showed that many of the predictions based on BLAST did not stand
298 up to hypothesis testing, and suggest that the similarity between sequences that cause high alien
299 indices do not necessarily provide true phylogenetic signal. Consequently, incorporation of
300 phylogenetic likelihood-based methods are necessary when performing HGT analyses in animals.

301 The importance of HGT as an evolutionary mechanism in prokaryotes and eukaryotes has
302 been underestimated. While studies of HGT in animals are gradually becoming more accepted,
303 many challenges remain to quantify the extent of HGT and its impacts. To mitigate some of these
304 challenges, rigorous approaches that employ both BLAST- and phylogenetic likelihood-based

methods should be applied to future HGT studies in animals. Here we provided evidence of HGT in ctenophores by applying these rigorous methods, and found similar patterns of transfer between prokaryotes and eukaryotes with preference for operational genes. However, many more studies will be necessary to gain a comprehensive overview of HGT and the mechanisms by which HGT occurs in animals.

Acknowledgements

This work was supported by startup funds to J.F.R. from the University of Florida DSP Research Strategic Initiatives and the Office of the Provost. This material is based upon work supported by the National Science Foundation under Grant No. (1542597). A.M.H. received support from the McKnight Doctoral Fellowship Program.

References

Alamaru, A., Brokovich, E., & Loya, Y. (2015). Four new species and three new records of benthic ctenophores (Family: Coeloplanidae) from the Red Sea. *Marine Biodiversity,* 46(1): 261-279. https://doi.org/10.1007/s12526-015-0362-4

Andersson, J. O., et al. (2001). Are There Bugs in Our Genome? *Science*, 292(5523): 1848-1850.

Andersson, J. O., et al. (2006). Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes. *BMC Evolutionary Biology*, 6(1): 27. https://doi.org/10.1186/1471-2148-6-27

Arcila, D., et al. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(20). https://doi.org/10.1038/s41559-016-0020

Artamonova, I. I., & Mushegian, A. R. (2013). Genome sequence analysis indicates that the model eukaryote Nematostella vectensis harbors bacterial consorts. *Applied and Environmental Microbiology*, 79(22): 6868–73. https://doi.org/10.1128/AEM.01635-13

Artamonova, I. I., et al. (2015). Prokaryotic genes in eukaryotic genome sequences: when to infer horizontal gene transfer and when to suspect an actual microbe. *Environmental Microbiology*, 17(7): 2203–2208. https://doi.org/10.1111/1462-2920.12854

Bapteste E., Moreira D., Philippe H. (2003). Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. *Gene*, 318: 185-91.

Bergthorsson, U., et al. (2004). Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm Amborella. *Proceedings of the National Academy of Sciences of the United States of America,* 101(51): 17747-17752. https://doi.org/10.1073/pnas.0408336102

Bhattacharya, D., et al. (2013). Genome Analysis of Elysia chlorotica Egg DNA Provides No Evidence for Horizontal Gene Transfer into the Germ Line of This Kleptoplastic Mollusc. *Molecular Biology and Evolution*, 30(8): 1843–1852. https://doi.org/10.1093/molbev/mst084

Boothby T. C., et al. (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52):15976-81.

Borowiec, M. L., et al. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics*, 16(987): 987. https://doi.org/10.1186/s12864-015-2146-4

348 Boschetti, C., et al. (2012). Biochemical Diversification through Foreign Gene Expression in
349 Bdelloid Rotifers. *PLoS Genetics*, 8(11): e1003035.
350 https://doi.org/10.1371/journal.pgen.1003035

351 Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. *Proceedings.*
352 *Biological Sciences*, 277(1683): 819–27. https://doi.org/10.1098/rspb.2009.1679

353 Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans.
354 *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1777): 20132450.
355 http://dx.doi.org/10.1098/rspb.2013.2450

356 Buss, L. W. (1983). Evolution, development, and the units of selection. *Proceedings of the*
357 *National Academy of Sciences*, 80(5): 1387-1391.

358 Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in
359 phylogenetic analysis. *Molecular Biology and Evolution,* 17(4): 540-552.
360 https://doi.org/10.1093/oxfordjournals.molbev.a026334

361 Chang, E. S., et al. (2015). Genomic insights into the evolutionary origin of Myxozoa within
362 Cnidaria. *Proceedings of the National Academy of Sciences of the United States of America*,
363 112(48): 14912–7. https://doi.org/10.1073/pnas.1511468112

364 Chapman, J. A., et al. (2010). The dynamic genome of Hydra. *Nature*, 464(7288): 592–596.
365 https://doi.org/10.1038/nature08830

366 Church, S. H., Ryan, J. F., & Dunn, C. W. (2015). Automation and Evaluation of the SOWH Test
367 with SOWHAT. *Systematic Biology*, 64(6): 1048–58. https://doi.org/10.1093/sysbio/syv055

368 Daniels, C., & Breitbart, M. (2012). Bacterial communities associated with the ctenophores
369 Mnemiopsis leidyi and Beroe ovata. *FEMS Microbiology Ecology*, 82(1): 90–101.
370 https://doi.org/10.1111/j.1574-6941.2012.01409.x

371 Dehal, P., et al. (2002). The Draft Genome of Ciona intestinalis: Insights into Chordate and
372 Vertebrate Origins. *Science,* 298(5601): 2157-67. https://doi.org/10.1126/science.1080049

373 Delmont, T. O., & Eren, A. M. (2016). Identifying contamination with advanced visualization and
374 analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, *4*,
375 e1839. https://doi.org/10.7717/peerj.1839

376 Doolittle, W. F. (1999). Phylogenetic Classification and the Universal Tree. *Science*, 284(5423):
377 2124-2128.

378 Dunn, C. W., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree
379 of life. *Nature*, 452(7188): 745–749. https://doi.org/10.1038/nature06614

380 Eliáš, M., et al. (2016). A paneukaryotic genomic analysis of the small GTPase RABL2
381 underscores the significance of recurrent gene loss in eukaryote evolution. *Biology direct*,
382 11(1): 5. https://doi.org/10.1186/s13062-016-0107-8

383 Fraune, S., & Bosch, T. C. G. (2007). Long-term maintenance of species-specific bacterial
384 microbiota in the basal metazoan Hydra. *Proceedings of the National Academy of Sciences*
385 *of the United States of America*, 104(32): 13146–13151.
386 https://doi.org/10.1073/PNAS.0703375104

387 Garcia-Vallvé, S., Romeu, A., & Palau, J. (2000). Horizontal gene transfer in bacterial and
388 archaeal complete genomes. *Genome Research*, 10(11): 1719–25.
389 https://doi.org/10.1101/GR.130000

390 Gladyshev, E. A., Meselson, M., & Arkhipova, I. R. (2008). Massive horizontal gene transfer in
391 bdelloid rotifers. *Science*, 320(5880): 1210–3. https://doi.org/10.1126/science.1156407

392 Glynn, P. W., et al. (2017). Benthic ctenophores (Platyctenida: Coeloplanidae) in south Florida:
393 environmental conditions, habitats, abundances, and behaviors. *Invertebrate Biology*.
394 136(4): 379-393. https://doi.org/10.1111/ivb.12189

395 Hall, C., Brachat, S., & Dietrich, F. S. (2005). Contribution of horizontal gene transfer to the
396 evolution of Saccharomyces cerevisiae. *Eukaryotic Cell*, 4(6): 1102–15.
397 https://doi.org/10.1128/EC.4.6.1102-1115.2005

398 Hejnol, A., et al. (2009). Assessing the root of bilaterian animals with scalable phylogenomic
399 methods. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1677): 4261–
400 4270. https://doi.org/10.1098/rspb.2009.0896

401 Hontzeas, N., et al. (2005). Evidence for horizontal transfer of 1-aminocyclopropane-1-
402 carboxylate deaminase genes. *Applied and Environmental Microbiology*, *71*(11), 7556–8.
403 https://doi.org/10.1128/AEM.71.11.7556-7558.2005

404 Jensen, L., et al. (2016). Assessing the effects of a sequestered germline on interdomain lateral
405 gene transfer in Metazoa. *Evolution*, 70(6): 1322–1333. https://doi.org/10.1111/evo.12935

406 Katoh, K., et al. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on
407 fast Fourier transform. *Nucleic Acids Research*, 30(14): 3059–3066.
408 https://doi.org/10.1093/nar/gkf436

409 Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
410 Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4): 772–
411 780. https://doi.org/10.1093/molbev/mst010

412 Katz, L. A. (2002). Lateral gene transfers and the evolution of eukaryotes: theories and data.
413 *International Journal of Systematic and Evolutionary Microbiology*, 52: 1893–1900.
414 https://doi.org/10.1099/00207713-52-5-1893

415 Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature*
416 *Reviews Genetics*, 9(8): 605–618. https://doi.org/10.1038/nrg2386

417 Koutsovoulos, G., et al. (2016). No evidence for extensive horizontal gene transfer in the genome
418 of the tardigrade Hypsibius dujardini. *Proceedings of the National Academy of Sciences of*
419 *the United States of America*, 113(18): 5053-5058. https://doi.org/10.1073/pnas.1600338113

Lander, E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921. https://doi.org/10.1038/35057062

Lawrence, J. G., & Roth, J. R. (1996). Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. *Genetics*, 143(4): 1843-1860.

Levin, M., et al. (2016). The mid-developmental transition and the evolution of animal body plans. *Nature*, 531(7596): 637-41. https://doi.org/10.1038/nature16994

Longo, M. S., et al. (2011). Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. *PLoS ONE*, 6(2): e16410. https://doi.org/10.1371/journal.pone.0016410

Madhusoodanan J. (2015) Horizontal Gene Transfer a Hallmark of Animal Genomes? The Scientist, http://www.the-scientist.com/?articles.view/articleNo/42420/title/Horizontal-Gene-Transfer-a-Hallmark-of-Animal-Genomes-/

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, 2: e675. https://doi.org/10.7717/peerj.675

Moreland, R. T., et al. (2014). A customized Web portal for the genome of the ctenophore Mnemiopsis leidyi. *BMC genomics*, 15(1): 316.

Moroz, L. L., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, 510(7503): 109–114. https://doi.org/10.1038/nature13400

Moss, A. G., et al. (2001). Protistan epibionts of the ctenophore Mnemiopsis mccradyi Mayer. *Hydrobiologia*, 451: 295–304.

Naccache, S. N., et al. (2013). The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of Virology*, 87(22): 11966–77. https://doi.org/10.1128/JVI.02323-13

O'Leary N. A., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1): D733-45.

Paradis E., Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics,* 20: 289-290.

Pible, O., et al. (2014). The importance of recognizing and reporting sequence database contamination for proteomics. *EuPA Open Proteomics*, 3: 246-249. https://doi.org/10.1016/j.euprot.2014.04.001

Pisani, D., et al. (2015). Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of Sciences*, 112(50): 15402-15407. https://doi.org/10.1073/pnas.1518127112

Pruitt, K.D., Tatusova, T., Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research,* 35(Database issue): D61-65. http://doi.org/10.1093/nar/gkl842

Radzvilavicius, A. L., et al. (2016). Selection for Mitochondrial Quality Drives Evolution of the Germline. *PLoS Biology*, 14(12): e2000410. https://doi.org/10.1371/journal.pbio.2000410

Richards T. A., et al. (2006). Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryotic Cell,* 5(9): 1517-31. https://dx.doi.org/10.1128%2FEC.00106-06

Ryan, J. F., et al. (2013). The Genome of the Ctenophore *Mnemiopsis leidyi* and Its Implications for Cell Type Evolution. *Science*, 342(6164): 1242592. https://doi.org/10.1126/science.1242592

Salter, S. J., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology,* 12(1): 87. https://doi.org/10.1186/s12915-014-0087-z

Salzberg, S. L., et al. (2001). Microbial Genes in the Human Genome: Lateral Transfer or Gene Loss? *Science*, 292(5523): 1903-1906. https://doi.org/10.1126/science.1061036

Schönknecht, G., Weber, A. P. M., & Lercher, M. J. (2014). Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *BioEssays*, 36(1): 9–20. https://doi.org/10.1002/bies.201300095

Shen, X.-X., Todd Hittinger, C., & Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Publishing Group*, 1(0126). https://doi.org/10.1038/s41559-017-0126

Shimodaira, H., & Hasegawa, M. (1999). Letter to the Editor Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, 16(8): 1114–1116.

Shimodaira, H., & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *BIOINFORMATICS APPLICATIONS NOTE*, 17(12): 1246–1247.

Simion, P., et al. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, 27(7): 958-967. http://dx.doi.org/10.1016/j.cub.2017.02.031

Song, J.-I., & Hwang, S.-J. (2010). New Species of Genus Coeloplana (Ctenophora: Tentaculata: Platyctenida) from Korea. *Animal Systematics, Evolution and Diversity*, 26(3): 217–221. https://doi.org/10.5635/KJSZ.2010.26.3.217

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8): 472–482. https://doi.org/10.1038/nrg3962

Stamatakis, A. (2014) RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30: 1312–1313.

Stanhope, M. J., et al. (2001). Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 411(6840): 940–944. https://doi.org/10.1038/35082058

Strong, M. J., et al. (2014). Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens*, 10(11): e1004437. https://doi.org/10.1371/journal.ppat.1004437

Swofford, D.L., et al. (1996). Phylogenetic inference. Molecular systematics. 2nd ed. Sunderland, MA: Sinauer, 407–514.

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, *3*(9), 711–721. https://doi.org/10.1038/nrmicro1234

Torruella, G., et al. (2015). Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology*, 25(18): 2404–2410. https://doi.org/10.1016/j.cub.2015.07.053

Weinert, L. A., Welch, J. J., & Jiggins, F. M. (2009). Conjugation genes are common throughout the genus Rickettsia and are transmitted horizontally. *Proceedings of the Royal Society of London B: Biological Sciences,* 276(1673): 3619–3627. https://doi.org/10.1098/rspb.2009.0875

Whelan, N. V., et al. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences of the United States of America*, 112(18): 5773–5778. https://doi.org/10.1073/pnas.1503453112

Whelan, N. V., et al. (2017). Ctenophore relationships and their placement as the sister group to all other animals. *Nature Ecology and Evolution*, 1: 1737–1746. http://doi.org/10.1038/s41559-017-0331-3

Willerslev, E., et al. (2002). Contamination in the Draft of the Human Genome Masquerades As Lateral Gene Transfer. *DNA Sequence,* 13(2): 75-76. https://doi.org/10.1080/10425170290023392

Zhu, B., et al. (2011) Horizontal gene transfer in silkworm, Bombyx mori. *BMC Genomics*, 12(1): 248. https://doi.org/10.1186/1471-2164-12-248

Figure 1. Comparisons of confirmed HGTs identified in the *M. leidyi* genome to other ctenophore species' transcriptomes. Purple boxes indicate the specified HGT is present in the

519 species' transcriptome confirmed by reciprocal best BLAST hits; white boxes indicate the gene is
520 absent in the species' transcriptome. Tree was inferred by Moroz et al. (2014).

521 **Figure 2. Maximum-likelihood best tree and metazoan-constraint tree compared in the**
522 **SOWH and AU tests.** Gene IDs (in black) denote the putative HGTs. **(**A), (C), and (F) are
523 examples of RAxML best trees for HGTs confirmed by phylogenetic analyses and hypothesis
524 testing. (B), (D), and (F) are examples of trees where putative HGTs have been constrained to
525 produce monophyletic Animalia and have been optimized in RAxML. Taxa that are prefixed
526 "META_" are from our alien_index database version 0.01 (i.e., META_NVEC (*Nematostella*
527 *vectensis*), META_TADH (*Trichoplax adhaerens*), META_HSAP (*Homo sapiens*), META_CTEL
528 (*Capitella teleta*), META_DMEL (*Drosophila melanogaster*), META_AQUE (*Amphimedon*
529 *queenslandica*). MET=Metazoa; BAC=Bacteria; EUK=Eukaryota; FUN=Fungi; More details for
530 each taxon are specified in Table S3.

531 **Table 1. Hypothesis testing on HGT candidates that were confirmed by phylogenetic**
532 **analyses**. P-values indicate the level of support for HGTs in comparison to the metazoan
533 constraint tree for the SOWH test and suboptimal trees (bootstrap and manually generated) in the
534 AU test. Candidates in blue have significant values in all three tests ($p \leq 0.05$) and are likely
535 HGTs.

536 **Figure 3. A comparison of likelihood proportions between the best tree, metazoan-**
537 **constrained tree, and bootstrap trees for HGT candidates with BLAST hits to Metazoa.**
538 Likelihood proportions are individual likelihood values divided by the average likelihood value

539 for suboptimal trees (i.e. bootstrap trees). Red points indicate likelihood proportions of the best
540 tree (i.e., tree indicating HGT). Blue points indicate likelihood proportions of the metazoan
541 constrained tree (i.e., tree contradicting HGT). The violin plot shows the distribution of
542 likelihood proportions of 100 bootstrap trees for each HGT candidate. Confirmed HGTs were
543 validated by phylogenetic analyses and hypothesis testing.

544 **Figure 4. Expression profiles of 6 likely HGTs.** Single-embryo RNA-Seq analyses were
545 performed over 20 hours. Confirmed HGTs with transcript-per-million values (medians for each
546 set of time-point replicates) greater than or equal to 100 over 20 hours (25 time points) are
547 shown.

548 **Table 2. HGT origins and functions.** HGT functions were determined by BLAST against the
549 UniProt database and associated Pfam-A domains were searched on the *Mnemiopsis* Genome
550 Portal. The origin column shows the domains of life from which these genes are predicted to have
551 been transferred (Bac = Bacteria; Euk = Eukaryota). The RefSeq column shows a more detailed
552 classification for the origin of HGTs. All rows highlighted in orange indicate genes that show
553 early developmental expression.

Figure 1(on next page)

Comparisons of confirmed HGTs identified in the *M. leidyi* genome to other ctenophore species' transcriptomes.

Purple boxes indicate the specified HGT is present in the species' transcriptome confirmed by reciprocal best BLAST hits; white boxes indicate the gene is absent in the species' transcriptome. Tree was inferred by Moroz et al. (2014).

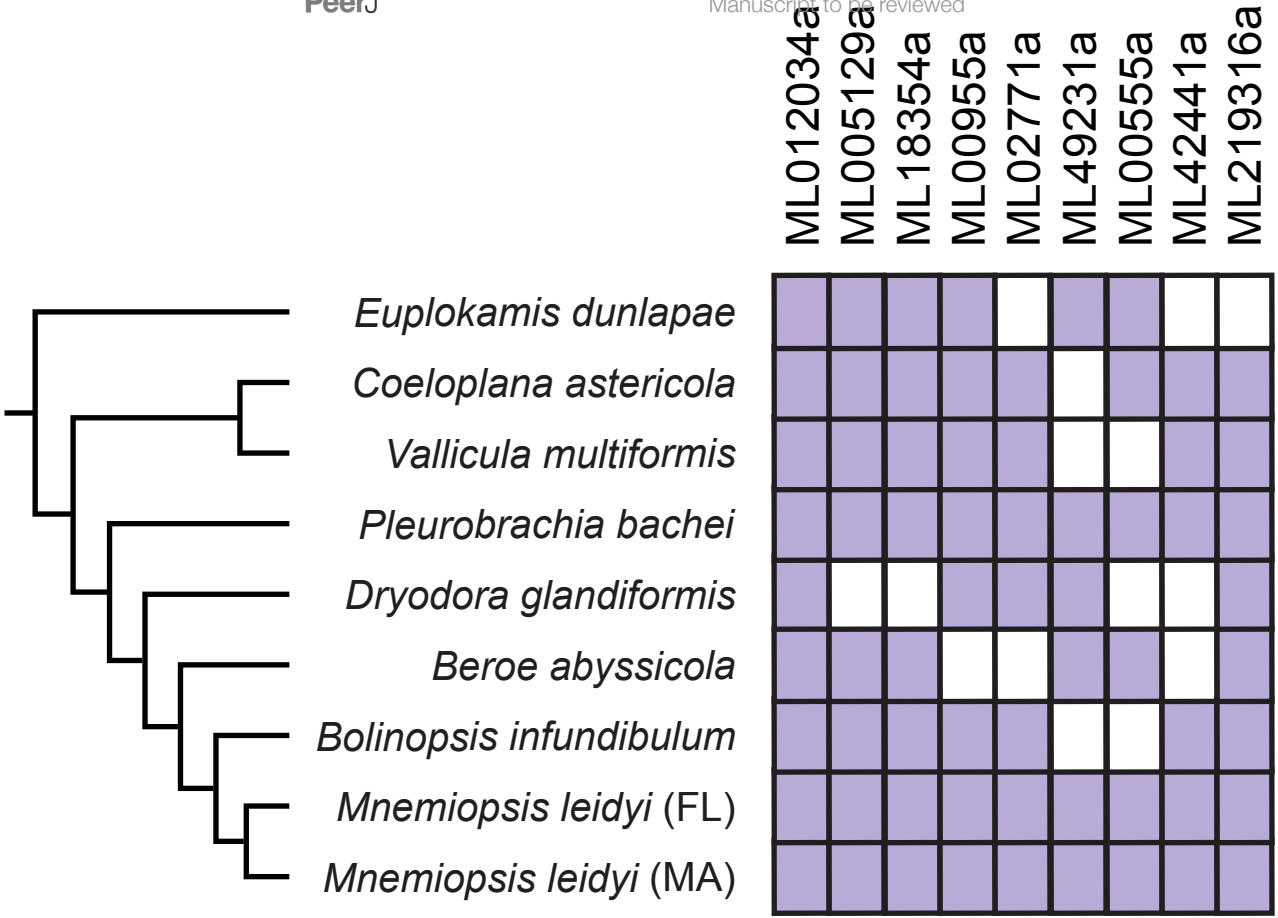
ML012034a
ML005129a
ML18354a
ML00955a
ML02771a
ML492231a
ML00555a
ML42441a
ML219316a
Euplokamis dunlapae
Coeloplana astericola
Vallicula multiformis
Pleurobrachia bachei
Dryodora glandiformis
Beroe abyssicola
Bolinopsis infundibulum
Mnemiopsis leidyi (FL)
Mnemiopsis leidyi (MA)

Figure 2(on next page)

Maximum-likelihood best tree and metazoan-constraint tree compared in the SOWH and AU tests.

Gene IDs (in black) denote the putative HGTs. (A), (C), and (F) are examples of RAxML best trees for HGTs confirmed by phylogenetic analyses and hypothesis testing. (B), (D), and (F) are examples of trees where putative HGTs have been constrained to produce monophyletic Animalia and have been optimized in RAxML. Taxa that are prefixed “META_” are from our alien_index database version 0.01 (i.e., META_NVEC (Nematostella vectensis), META_TADH (Trichoplax adhaerens), META_HSAP (Homo sapiens), META_CTEL (Capitella teleta), META_DMEL (Drosophila melanogaster), META_AQUE (Amphimedon queenslandica). MET=Metazoa; BAC=Bacteria; EUK=Eukaryota; FUN=Fungi; More details for each taxon are specified in Table S3.

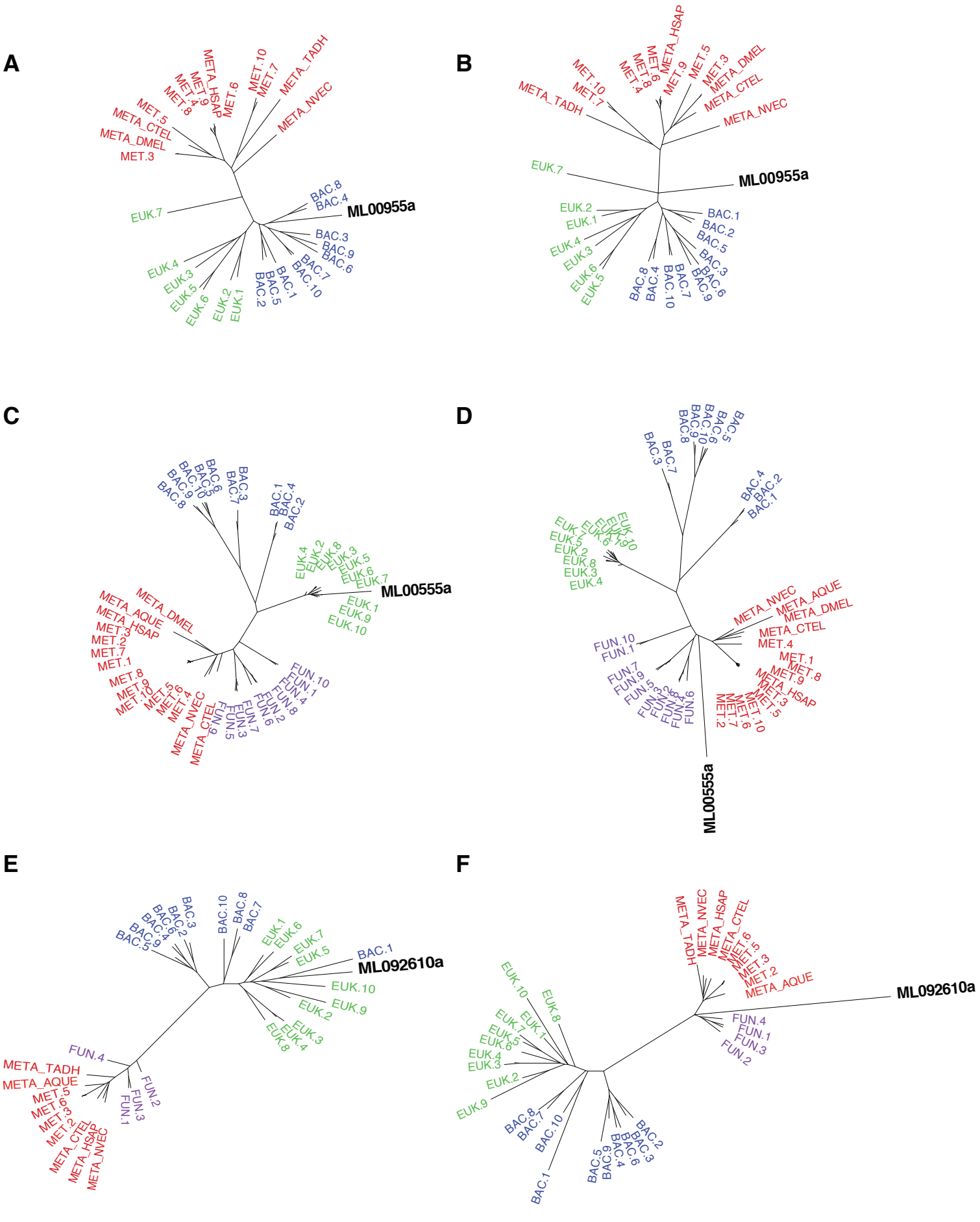
A
META_HSAP
MET.9
MET.4
MET.8
MET.6
MET.10
MET.7
META_TADH
MET.5
META_CTEL
META_DMEL
MET.3
META_NVEC
EUK.7
BAC.8
BAC.4
ML00955a
BAC.3
BAC.9
BAC.6
BAC.7
BAC.10
BAC.1
BAC.5
BAC.2
EUK.4
EUK.3
EUK.5
EUK.6
EUK.2
EUK.1
B
MET.10
MET.7
META_TADH
META_HSAP
MET.6
MET.8
MET.4
MET.9
MET.5
MET.3
META_DMEL
META_CTEL
META_NVEC
EUK.7
ML00955a
EUK.2
EUK.1
EUK.4
EUK.3
EUK.6
EUK.5
BAC.1
BAC.2
BAC.5
BAC.3
BAC.6
BAC.9
BAC.8
BAC.4
BAC.10
BAC.7
C
BAC.6
BAC.5
BAC.10
BAC.9
BAC.8
BAC.3
BAC.7
BAC.1
BAC.4
BAC.2
EUK.4
EUK.2
EUK.8
EUK.3
EUK.5
EUK.6
EUK.7
ML00555a
EUK.1
EUK.9
EUK.10
META_DMEL
META_AQUE
META_HSAP
MET.3
MET.2
MET.7
MET.1
MET.8
MET.9
MET.10
MET.5
MET.6
MET.4
META_NVEC
META_CTEL
FUN.9
FUN.5
FUN.3
FUN.7
FUN.2
FUN.6
FUN.8
FUN.4
FUN.1
FUN.10
D
BAC.5
BAC.6
BAC.10
BAC.9
BAC.8
BAC.3
BAC.7
BAC.4
BAC.2
BAC.1
EUK.5
EUK.6
EUK.10
EUK.9
EUK.2
EUK.8
EUK.3
EUK.4
META_NVEC
META_AQUE
META_DMEL
META_CTEL
MET.4
FUN.10
FUN.1
MET.1
MET.8
MET.9
META_HSAP
MET.3
MET.5
MET.10
MET.6
MET.7
MET.2
FUN.7
FUN.9
FUN.5
FUN.3
FUN.2
FUN.8
FUN.4
FUN.6
ML00555a
E
BAC.3
BAC.10
BAC.8
BAC.7
BAC.6
BAC.2
BAC.4
BAC.9
BAC.5
EUK.1
EUK.6
EUK.7
EUK.5
BAC.1
ML092610a
EUK.10
EUK.9
EUK.2
EUK.3
EUK.4
EUK.8
FUN.4
META_TADH
META_AQUE
MET.5
MET.6
MET.3
MET.7
MET.2
META_CTEL
META_HSAP
META_NVEC
FUN.2
FUN.3
FUN.1
F
META_TADH
META_NVEC
META_HSAP
META_CTEL
MET.6
MET.5
MET.3
MET.2
META_AQUE
ML092610a
FUN.4
FUN.1
FUN.3
FUN.2
EUK.10
EUK.8
EUK.1
EUK.7
EUK.5
EUK.6
EUK.4
EUK.3
EUK.2
EUK.9
BAC.8
BAC.7
BAC.10
BAC.1
BAC.5
BAC.9
BAC.4
BAC.6
BAC.3
BAC.2

Table 1(on next page)

Hypothesis testing on HGT candidates that were confirmed by phylogenetic analyses.

P-values indicate the level of support for HGTs in comparison to the metazoan constraint tree for the SOWH test and suboptimal trees (bootstrap and manually generated) in the AU test. Candidates in blue have significant values in all three tests (p 0.05) and are likely HGTs.

Genes	SOWH p-value	AU Bootstrap p-value	AU Manual p-value
ML00555a	<0.001	4.00E-45	7.00E-06
ML49231a	<0.001	2.00E-44	7.00E-103
ML092610a	<0.001	2.00E-31	4.00E-68
ML005129a	<0.001	1.00E-04	6.00E-06
ML00955a	<0.001	0.021	0.002
ML02771a	<0.001	0.023	0.029
ML42441a	<0.001	0.047	0.022
ML177319a	<0.001	0.226	0.042
ML120721a	<0.001	0.48	0.245
ML049014a	0.985	0.862	0.604
ML070218a	0.262	0.849	0.361
ML102910a	0.229	0.719	0.255

Figure 3(on next page)

A comparison of likelihood proportions between the best tree, metazoan-constrained tree, and bootstrap trees for HGT candidates with BLAST hits to Metazoa.

Likelihood proportions are individual likelihood values divided by the average likelihood value for suboptimal trees (i.e. bootstrap trees). Red points indicate likelihood proportions of the best tree (i.e., tree indicating HGT). Blue points indicate likelihood proportions of the metazoan constrained tree (i.e., tree contradicting HGT). The violin plot shows the distribution of likelihood proportions of 100 bootstrap trees for each HGT candidate. Confirmed HGTs were validated by phylogenetic analyses and hypothesis testing.

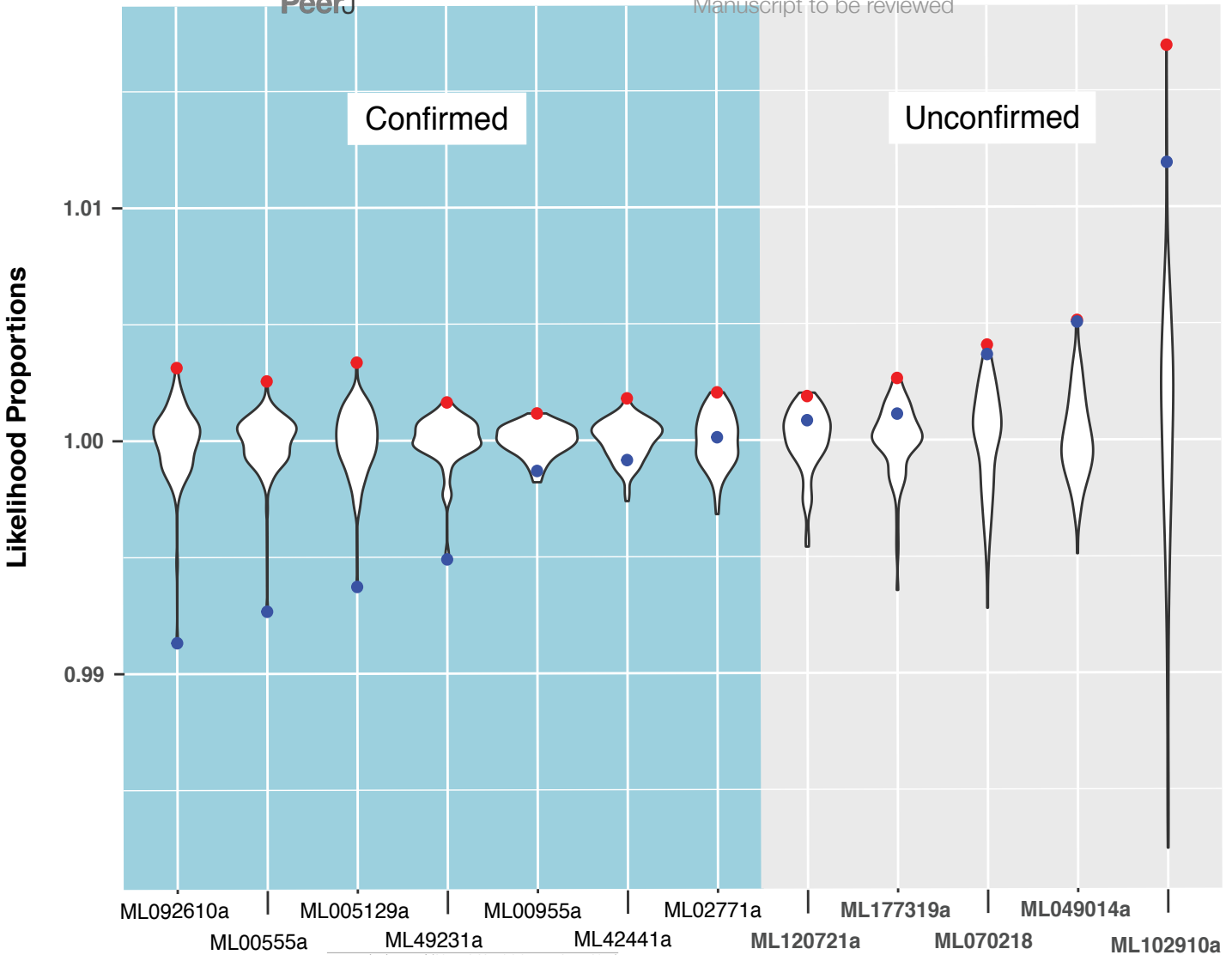
Confirmed
Unconfirmed
Likelihood Proportions
1.01
1.00
0.99
ML092610a
ML00555a
ML005129a
ML49231a
ML00955a
ML42441a
ML02771a
ML120721a
ML177319a
ML070218
ML049014a
ML102910a

Figure 4(on next page)

Expression profiles of 6 likely HGTs.

Single-embryo RNA-Seq analyses were performed over 20 hours. Confirmed HGTs with transcript-per-million values (medians for each set of time-point replicates) greater than or equal to 100 over 20 hours (25 time points) are shown.

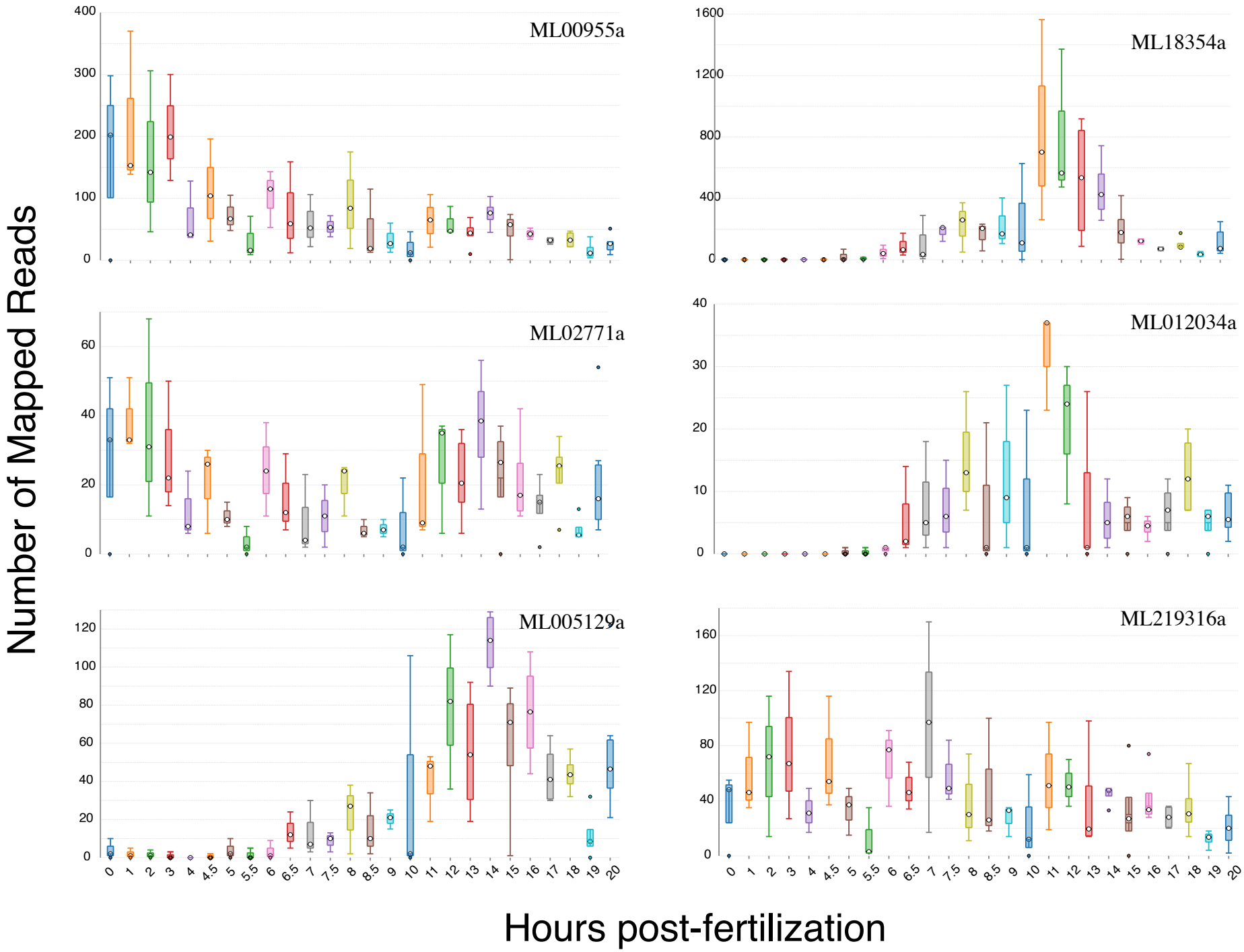

Table 2(on next page)

HGT origins and functions.

HGT functions were determined by BLAST against the UniProt database and associated Pfam-A domains were searched on the *Mnemiopsis* Genome Portal. The origin column shows the domains of life from which these genes are predicted to have been transferred (Bac = Bacteria; Euk = Eukaryota). The RefSeq column shows a more detailed classification for the origin of HGTs. All rows highlighted in orange indicate genes that show early developmental expression.

Genes	Function	Domains	Origin	RefSeq
ML00955a	Putative metalloendopeptidase	Peptidase family M13	Bac	Proteobacteria
ML005129a	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein	2OG-Fe(II) oxygenase superfamily	Euk	Stramenopiles
ML02771a	Penicillin acylase	Penicillin amidase	Bac	Proteobacteria
ML012034a	Uncharacterized protein	2OG-Fe(II) oxygenase superfamily	Euk	Stramenopiles
ML18354a	Putative chalcone and stilbene synthase	Chalcone and stilbene synthases, 3-Oxoacyl- synthase III, FAE1/Type III polyketide synthase	Bac	Bacteroidetes
ML219316a	Uncharacterized protein		Bac	Planctomycetes
ML00555a	Phospholipase D alpha 1	C2, Phospholipase D	Euk	Viridiplantae
ML49231a	Phospholipase D gamma 1	Phospholipase D	Euk	Rhodophyta
ML42441a	NADH dehydrogenase, putative	Pyridine nucleotide-disulphide oxidoreductase	Euk	Amoebozoa