Horizontally transferred genes in the ctenophore *Mnemiopsis leidyi*

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) are homologous to characterized enzymes, supporting previous observations that genes encoding enzymes are more likely to be retained after HGT events. We found that the majority of these nine horizontally transferred genes were expressed during 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.

1   **Horizontally transferred genes in the ctenophore *Mnemiopsis leidyi***

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

5   [1] Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL, USA

6   [2] Department of Biology, University of Florida, Gainesville, FL, USA

8   Corresponding Author:

9   Joseph Ryan[1]

10   joseph.ryan@whitney.ufl.edu

11 **Abstract**

13 Horizontal gene transfer has had major impacts on the biology of a wide range of organisms

14 from antibiotic resistance in bacteria to adaptations to herbivory in arthropods. A growing body

15 of literature shows that horizontal gene transfer (HGT) between non-animals and animals is more

16 commonplace than previously thought. In this study, we present a thorough investigation of HGT

17 in the ctenophore *Mnemiopsis leidyi*. We applied tests of phylogenetic incongruence to identify

18 nine genes that were likely transferred horizontally early in ctenophore evolution from bacteria

19 and non-metazoan eukaryotes. All but one of these HGTs (an uncharacterized protein) are

20 homologous to characterized enzymes, supporting previous observations that genes encoding

21 enzymes are more likely to be retained after HGT events. We found that the majority of these

22 nine horizontally transferred genes were expressed during development, suggesting that they are

23 active and play a role in the biology of *M. leidyi*. This is the first report of HGT in ctenophores,

24 and contributes to an ever-growing literature on the prevalence of genetic information flowing

25 between non-animals and animals.

26   **Introduction**

28   Evolution is commonly thought to occur by descent with modification from a single

29   lineage. However, evidence has shown that genomes from bacteria, archaea, and eukaryotes are

30   typically chimeric, resulting from horizontal (or lateral) gene transfers (Garcia-Vallvé et al.

31   2000; Katz 2002). As such, horizontal gene transfer (HGT) has likely impacted evolution more

32   than originally thought by creating opportunities for rapid genetic diversification and

33   contributing to speciation events. Moreover, HGT is a potential catalyst for organisms to acquire

34   novel traits (Soucy et al. 2015) and creates opportunities for HGT receivers to exploit new

35   ecological niches (Boto 2010). For example, HGTs have played an important role in herbivory in

36   arthropods (Wybouw et al. 2016), venom recruitment in parasitoid wasps (Martinson et al.

37   2016), cellulose production in urochordates (Dehal et al. 2002) and plant parasitism in

38   nematodes (Haegeman et al. 2011).

40   Although HGT is generally accepted as an important evolutionary mechanism in

41   prokaryotes (Boto 2014), it remains controversial whether it occurs in animals, despite many

42   convincing studies (Madhusoodanan 2015). Much of the skepticism has been fueled by high-

43   profile reports of HGT (e.g., Lander et al. 2001; Boothby et al. 2015) that were later shown to be

44   largely incorrect due to contamination or taxon sampling (Stanhope et al. 2001; Koutsovoulos et

45   al. 2016). In addition, HGT in animals is hypothesized to be rare due to the origin of a

46   sequestered germ line, which provides fewer opportunities for germ cells to be exposed to

47   foreign DNA (Doolittle 1999; Andersson et al. 2001; Jensen et al. 2016). However, the presence

48   and absence of germline sequestration is not well described across the animal tree of life, and

49   there are inconsistencies between studies regarding which animal groups have sequestered

50   germlines (Buss, 1983; Radzvilavicius et al. 2016; Jensen et al. 2016).

51

52   The major challenges for HGT detection efforts have been taxon sampling and

53   contamination. Many early reports of HGT in animals were overturned due to limited

54   representation of taxa in public genomic databases (e.g., Salzberg et al. 2001). For example, a

55   gene present in bacteria and humans, but absent from nematodes and drosophilids (the most

56   highly represented taxa at the time) may have been considered the result of HGT, until

57   discovering that the gene is present in many other animal genomes that were not available at the

58   time of the initial claim. In these cases, the limited representation of taxa made it difficult to

59   distinguish HGTs from differential gene loss (Andersson et al. 2006; Keeling & Palmer, 2008).

60   More recently, contamination has led to both overestimation and likely underestimation of HGT

61   events. In several recent cases, contamination in newly generated datasets has been interpreted as

62   HGT but later shown to be cross-contaminants present in genome sequences (Bhattacharya et al.

63   2013; Delmont & Eren 2016; Koutsovoulos et al. 2016). On the other hand, the presence of

64   contaminants in public databases (e.g., a bacteria sequence labeled as an animal sequence) makes

65   it difficult to identify *bona fide* HGTs, as "animal" sequences will appear among the top BLAST

66   hits for a particular HGT, leading to false negatives. As such, contamination remains a major

67   hurdle to contemporary studies of HGT.

68

69   Pairwise BLAST-based similarity scores (e.g., alien index (Gladyshev et al. 2008) and

70   the HGT index (Boschetti et al. 2012)) are the most common criteria used to detect HGT in

71   animals. However, these measures largely ignore phylogenetic information associated with

72   sequence data. While a positive BLAST-based result may be due to HGT, it may also result from

73   gene loss, selective evolutionary rates, convergent evolution, sequence contamination, and

74   species misassignment (Hall et al. 2005). Previous HGT studies have demonstrated that HGT

75   predictions need to be carefully considered and a combination of methods are required to rule out

76   false positives (Schönknecht et al. 2013). Hypothesis tests incorporating phylogenetic

77   incongruence are one such method that has been used to test HGT. While some studies in

78   animals have incorporated these techniques (e.g., Eliáš et al, 2016), they are more commonly

79   deployed in studies involving non-animals (e.g., Bapteste et al. 2003; Richards et al. 2006).

80

81   HGT has yet to be thoroughly explored in Ctenophora. Ctenophores (comb jellies) are

82   marine invertebrates that are morphologically characterized by eight rows of cilia used for

83   movement. They typically live in the water column, but the group includes benthic species as

84   well (Song & Hwang 2010; Alamaru et al. 2015; Glynn et al. 2017). Phylogenomic evidence

85   from studies including ctenophores has suggested that ctenophores are the sister group to all

86   other animals (Dunn et al. 2008; Hejnol et al. 2009; Ryan et al. 2013; Moroz et al. 2014;

87   Borowiec et al. 2015; Chang et al. 2015; Torruella et al. 2015; Whelan et al. 2015; Arcila et al.

88   2017; Shen et al. 2017; Whelan et al. 2017), but the position remains controversial with some

89   evidence supporting sponges as the sister group to the rest of animals (Philippe et al. 2009; Pick

90   et al. 2010; Pisani et al. 2015; Telford et al. 2015; Simion et al. 2017; Feuda et al. 2017). Thus,

91   investigating HGT in ctenophores is essential to understanding its implications on early animal

92   evolution.

93

94   Here, we apply a rigorous framework to identify and confirm HGTs in the ctenophore

95   *Mnemiopsis leidyi.* Our process includes identification of HGT candidates by alien index and

96   confirmation by phylogenetic hypothesis testing, providing statistical support in an evolutionary

97   framework. Furthermore, we analyze gene expression profiles during development to obtain

98   clues as to the function of these HGTs in *M. leidyi.*

99

100   **Material and Methods**

101

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

103

104   **Identification of HGT candidates by alien_index**

105

106   As part of this project, we developed the program alien_index and complimentary

107   metazoan/non-metazoan sequence databases to automate the generation of alien index

108   (Gladyshev et al. 2008) and HGT index scores (Boschetti et al. 2012). We BLASTed the entire

109   set of *M. leidyi* gene models (ML2.2) against a database of animal and non-animal sequences

110   (alien_index_db version 0.01) and then calculated alien index values as the logarithmic

111   difference between the best BLASTP E-values for animal and non-animal hits (as outlined in

112   Gladyshev et al. 2008) (Fig. 1A). In more simple terminology, the alien index reflects the

113   difference between the E-value of the best non-animal BLAST hit and that of the best animal hit.

114   The database used includes translated gene models from curated genomes that include bacteria

115   (5), archaea (2), non-animal eukaryotes (5), and animals (12). See Table S1 or

116   http://ryanlab.whitney.ufl.edu/downloads/alien_index/ for the entire list of taxa. HGT index

117   values were computed by the difference in the highest non-metazoan and metazoan bit scores

118   generated from the alien_index database. The alien_index program is available at:

119   https://github.com/josephryan/alien_index

120

121   **Confirmation of HGTs**

122

123   We applied a phylogenetic approach to confirm putative HGTs. HGT candidates from

124   alien_index were used as queries for BLASTP against NCBI's RefSeq database (O'Leary et al.

125   2016) using the NCBI BLAST interface. We collected the top ten sequences each from bacteria,

126   eukaryotes, fungi, and animals with an E-value cutoff of 0.1. We included only the first sequence

127   if there were hits to sequences from species in the same genus (Fig. 1B). We also added the top

128   BLAST hit (E-value ≤ 0.1) from each of the following fully sequenced animals *Amphimedon*

129   *queenslandica*, *Trichoplax adhaerens*, *Nematostella vectensis*, *Capitella teleta*, *Drosophila*

130   *melanogaster*, and *Homo sapiens* from version 0.01 of the alien_index database. Sequences were

131   aligned against the corresponding putative HGT using MAFFT (Katoh et al. 2002; Katoh &

132   Standley 2013) and trimmed with Gblockswrapper (Castresana 2000) (Fig. 1C). There were six

133   genes (ML012034a, ML06718a, ML03277a, ML02232a, ML18354a, ML219316a) with

134   BLASTP hits to non-animals but not to animals (E-value ≤ 0.1), preventing us from performing

135   additional phylogenetic analyses on these sequences. We considered the lack of animal BLASTP

136   hits below our cutoff as sufficient evidence that these six were clearly HGTs. ML018031a and

137   ML00882a only had two BLASTP hits to animal sequences. Since it was unclear if this resulted

138   from contamination, we were unable to test these genes using phylogenetic approaches, so they

139   were removed from contention as HGTs.

140

141   We performed maximum-likelihood analyses on the remaining 29 alignments using

142   RAxML version 8.1.21 (Stamatakis 2014) (Fig. 1C). Since the RefSeq database has many

143   instances of contamination (Pible et al. 2014), we allowed a maximum of two non-ctenophore

144   animal sequences to fall outside of the main animal clade. To implement this, we pruned putative

145   contaminants if the removal of two taxa resulted in a monophyletic animal clade (Fig. S1). We

146   discarded any HGT candidates with more than two taxa disrupting animal monophyly.

147

148   We explicitly tested topologies in opposition to HGT (i.e., animal monophyly) with the

149   SOWH test using SOWHAT (Church et al. 2015) and the AU test using CONSEL (Shimodaira

150   and Hasegawa 2001) (Fig. 1C). The SOWH and AU test evaluate statistical support for

151   phylogenetic incongruence by comparing the likelihood values between trees to a distribution of

152   trees generated by parametric sampling in the SOWH test and non-parametric sampling in the

153   AU test. We required that these two different approaches to hypothesis testing agreed to ensure

154   that our criteria confirming *bona fide* HGTs was stringent. To address any potential problems of

155   selection bias in the AU test (causing the likelihood value to bias upwards for the maximum

156   likelihood best tree when included in the dataset), we performed multiple AU analyses using

157   bootstrap trees as suboptimal trees (similar to Eliáš et al. 2016). We generated 100 bootstrap

158   trees using RAxML rapid bootstrap analyses, and verified there were no duplicate trees in our

159   100 bootstrap set using the ape package in R (Paradis et al. 2004). RAxML was used to generate

160   per-site log likelihoods for the best maximum-likelihood tree, the tree constraining the putative

161   HGT to metazoans (i.e., metazoan-constraint tree), and suboptimal trees, to perform the AU test

162   implemented through CONSEL. To test the effectiveness of comparing to bootstrap trees, we

163   manually created a set of suboptimal trees for each candidate HGT by shuffling clades of three

164   (Fig. S2) and running the same analyses. We evaluated the tree space covered by suboptimal

165   trees in the AU test (i.e., bootstrap and manually generated trees) by visualizing the data using

166   violin plots. We calculated likelihood proportions for each tree by dividing individual likelihood

167   scores by the average likelihood score of suboptimal trees. This was done to make the data

168   comparable for visualization since the likelihood scores differ between sets of gene trees. The

169   trees and scripts used to automate these phylogenetic analyses are available in the accompanying

170   GitHub site.

171   We verified that HGT candidates were not the result of bacterial contaminants by using

172   the *M. leidyi* genome browser (Moreland et al. 2014) to examine the intron/exon structure of

173   each HGT candidate, as well as the origin of their neighboring genes. We examined each intron

174   to determine whether they are actively handled by spliceosomes (which are only found in

175   eukaryotes), since bacteria, archaea, and viruses contain Group I and II introns. U2 spliceosomal

176   introns were identified by conserved GT dinucleotides at the 5' end and conserved AG

177   dinucleotides at the 3' end of introns. We also conducted reciprocal best BLASTP searches for

178   each HGT candidate (identified from the genome and gene models from an *M. leidyi* individual

179   collected in Woods Hole, MA) against the transcriptome of an *M. leidyi* individual collected

180   from St. Augustine, Florida, as well as seven other ctenophore transcriptomes reported in Moroz

181   et al. (2014): *Bolinopsis infundibulum, Beroe abyssicola, Dryodora glandiformis, Pleurobrachia*

182   *bachei, Vallicula multiformis, Coeloplana astericola, Euplokamis dunlapae.* For these searches

183   we used default parameters and an E-value cutoff of 0.1.

184

185   **HGT developmental expression profiles**

186

187   An extensive transcriptomic developmental timecourse of *M. leidyi* was recently

188   generated from single-embryos over the first 20 hours of embryogenesis (Levin et al. 2016). To

189   examine whether HGTs might play a role in development we used these data (GSE70185), as

190   well as additional time points for *M. leidyi* generated after this publication (GSE111748).

191

192   The Levin et al. (2016) data was produced by using three replicate timecourses that each

193   consisted of 20 isolated embryos from fertilization to 20 hours. Embryos were flash frozen and

194   RNA was extracted with TRIzol and sequenced using Illumina sequencing according to the

195   CEL-Seq protocol (Hashimshony et al. 2012). The reads were mapped to *M. leidyi* gene models

196   (ML2.2) using bowtie2 version 2.2.3 (Langmead & Salzberg 2012) with default settings and

197   counted using htseq-count (Anders et al. 2015). Since the CEL-Seq protocol considers only the

198   3' end of transcripts, the results occur in transcripts per million (tpm) without normalized

199   transcript lengths.

200

201   Since the publication of Levin et al. (2016), six additional time points (four replicates

202   each) for hours 14-19 (not included in the original study) have been sequenced and submitted to

203   the Gene Expression Omnibus (GSE111748). These additional data were produced by the same

204   researchers (i.e., Itai Yanai and Mark Martindale) from the original study using the same

205   methods and facilities. To create a baseline for what is considered adequate expression during

206   development, we summed median transcript-per-million values along the 25 time points for each

207   of our 9 confirmed HGTs. HGTs that had summed median read counts of 100 or greater were

208   classified as being expressed sufficiently to have roles in development. We chose a value that

209   was 10 times stricter than the minimum criteria in Levin et al. (2016) (i.e., 10 transcripts) to err

210   on the side of caution.

211

212   **HGT origins and functions**

213

214   To uncover the functional roles of HGTs, we used the BLAST interface provided by

215   UniProt and the UniProtKB database (Pundir et al. 2017) to identify homologous sequences to

216   characterize genes. Annotations of the top hits (E-value ≤0.1) were assigned to HGT candidates.

217   We also associated HGTs with Pfam-A domains using the MGP Portal under the *Mnemiopsis*

218   Gene Wiki (Moreland et al. 2014). In all cases, the annotations based on BLAST and Pfam-A

219   analyses were consistent with the results from our phylogenetic analyses.

220

221   **Results**

222

223   ***Mnemiopsis leidyi* HGTs**

224

225   Figure 1 shows our pipeline and results for each method during this analysis. We

226   calculated an alien index for every *M. leidyi* gene model using a database of 12 animals and 12

227   non-animals (Table S1). We identified 37 genes with alien indices greater than 45 and designated

228   these as HGT candidates (Fig. 1A; Table S2). In addition to the alien_index database, we

229   BLASTed the RefSeq database at NCBI restricting hits to bacteria, then to animals, and then to

230   non-animal eukaryotes (Fig. 1B). All but six HGT candidates had BLAST hits to animals with E-

231   values ≤ 0.1. We classified these six (ML012034a, ML06718a, ML03277a, ML02232a,

232   ML18354a, ML219316a) as absent from all other animals. We analyze these six further below.

233

234   For the remaining 29 candidates, we conducted detailed phylogenetic analyses using the

235   top 10 hits of unique non-animal and animal taxa from each of the RefSeq searches along with

236   sequences from *Amphimedon queenslandica*, *Trichoplax adhaerens*, *Nematostella vectensis*,

237   *Capitella teleta*, *Drosophila melanogaster*, and *Homo sapiens* that were top hits from our initial

238   BLASTs of the alien_index database (Fig. 1C). HGT candidates that formed a clade with all

239   other animals were ruled out as potential HGTs and candidates that disrupted animal monophyly

240   were tested further. We discarded 14 candidates with more than 2 non-ctenophore animal

241   sequences disrupting animal monophyly; in cases of 2 or less sequences, the disrupting

242   sequences were considered potential contaminants and pruned (e.g., Fig. S1). We discarded three

243   more candidates after pruning because the trees continued to result in a non-monophyletic clade

244   of animals. We then applied the SOWH and AU tests to the remaining 12 candidates to compare

245   the maximum-likelihood topology to the alternative hypothesis that candidate HGTs were more

246   closely related to animals (Fig. 2). This involved comparing likelihood values of optimal trees to

247   those that were constrained to produce a monophyletic Animalia. Our results showed that the AU

248   test was more conservative in confirming HGTs than the SOWH test (Table 1). For perspective

249   on how optimal trees compared to constrained trees, we ran AU tests comparing optimal and

250   constrained trees to bootstrap trees (Fig. 3). The likelihood scores of the constrained trees from

251   our confirmed HGTs in the AU test tend to fall outside or on the tails of the distribution of

252   likelihood scores of suboptimal trees, whereas the likelihood scores of constrained trees for

253   unconfirmed HGTs were all closer to the most likely tree than the bootstrap trees (Fig. 3). We

254   confirmed seven HGTs in which gene trees significantly differed (p < 0.05) from the metazoan

255   constraint trees in both the SOWH and AU analyses (Table 1).

256

257   We then analyzed the BLAST results of the seven HGT candidates confirmed by

258   phylogenetic analyses and the six genes which were absent in other animals (ML012034a,

259   ML06718a, ML03277a, ML02232a, ML18354a, ML219316a). We removed four of these genes

260   from contention (ML092610a, ML06718a, ML03277a, ML02232a) because the top hits from

261   BLAST against RefSeq were either Choanoflagellatea or Ichythosporea (the two non-animal

262   lineages most closely related to animals) (E-value ≤ 0.1). If ctenophores are the sister group to

263   the rest of animals, vertical inheritance remains a possibility for these cases. As such, these tests

264   support a total of nine HGTs.

265

266   We used the *M. leidyi* genome browser to examine the intron/exon structure of each of

267   these nine HGTs, as well as the origin of their neighboring genes for evidence of bacterial

268   contamination (lack of introns would indicate bacterial contamination). Eight HGTs were found

269   on scaffolds with intron-containing genes and eight HGTs contained introns (Table 2). The only

270   HGT not found on a scaffold with other intron-containing genes was ML49231a (itself

271   containing 6 introns). All intron-containing genes had U2 spliceosomal introns (except

272   ML219315a, a gene on the same scaffold as an HGT). These data suggest that our HGT

273   candidates were not bacterial contaminants.

274

275   To further test that these nine genes were not contaminants, we confirmed using BLAST

276   that these genes were also present in a transcriptome from an *M. leidyi* individual collected from

277 St. Augustine, FL (*M. leidyi* genome and gene models were from individuals collected in Woods

278 Hole, MA). We also performed reciprocal best BLASTP searches for each of the nine HGTs

279 against seven of the ctenophore transcriptomes published in Moroz et al. (2014): *Bolinopsis*

280 *infundibulum, Beroe abyssicola, Dryodora glandiformis, Pleurobrachia bachei, Vallicula*

281 *multiformis, Coeloplana astericola, Euplokamis dunlapae*. Each HGT was present in the

282 transcriptome of at least one other ctenophore species and in the Florida *M. leidyi* transcriptome

283 (Fig. 4). Because it is unlikely that the same species contaminated each of these datasets, these

284 comparisons provide additional evidence against HGT sequences resulting from contamination.

285 Here, we have included as much evidence as possible to carefully confirm nine HGTs in *M.*

286 *leidyi.*

287

288 **HGTs are expressed in development**

289

290   We summed transcript-per-million values (medians for each set of expression values at

291 25 time points) from single-embryo RNA-Seq analyses over 20 hours for each of the nine

292 confirmed HGTs to identify those HGTs that were expressed during development. Six of the

293 nine HGTs had sums greater than 100 (Fig. 5), suggesting that these had some role in

294 development. ML00955a was expressed maternally (0 hours post fertilization (hpf)) and

295 throughout early cleavage stages (1-3 hpf) with reduced expression later in development. Three

296 genes (ML005129a, ML18354a, ML012034a) were expressed later in development with spikes

297 during tentacle morphogenesis (9-12 hpf). ML02771a and ML219316a displayed cyclic

298 expression throughout development suggesting a potential role in cell cycle.

299

300   **HGTs are enzymes originating from non-animal eukaryotes and bacteria**

301

302   We used phylogenetic evidence to determine the origin of the nine HGTs. Four HGTs

303   originated from bacteria and five from non-animal eukaryotes (Table 3). We found no evidence

304   of HGTs that were transferred from Archaea. The five HGT events that originated from non-

305   animal eukaryotes appear to be transferred from Stramenopiles, Viriplantae, Rhodophyta or

306   Amoebozoa. To characterize gene function, we BLASTed the nine confirmed HGTs against the

307   UniProt database. All HGTs except one uncharacterized protein (ML219316a) were homologous

308   to known characterized enzymes (Table 3).

309

310   **Discussion**

311

312   **HGTs in ctenophores and their implications**

313

314   It had been previously speculated that ctenophores had HGTs since initial profiling

315   revealed that many 'bacteria-like' genes in ctenophores contained introns and were on

316   chromosomes with vertically inherited (i.e., non-HGT) genes (Artamonova et al. 2015). We

317   confirmed that all HGTs except ML012034a had spliceosomal introns and were on scaffolds

318   with other spliceosomal intron-containing genes (Table 2). This provided evidence that these

319   candidates were not the result of extrinsic contamination. We provided additional evidence that

320   candidates were not contaminants by showing that all HGTs were found in both Massachusetts

321   and Florida *M. leidyi* individuals, as well as in many other ctenophore species (Fig. 4). Six HGTs

322   were present in the *E. dunlapae* transcriptome suggesting that the majority of these HGT events

323   occurred very early in ctenophore evolution (Fig. 4). This deep evolutionary history suggests that

324   these HGTs may have had important impacts on the biology of ctenophores.

325

326   **Mechanisms driving HGT in ctenophores**

327

328   While we are uncertain about the mechanisms driving HGT, we speculate that some of

329   these may have resulted from symbiotic relationships with bacteria and non-animal eukaryotes.

330   *Gammaproteobacteria* and *Bacteroidetes* have been identified as two of the most abundant

331   groups of bacteria associated with ctenophores (Daniels & Breitbart 2012). These groups were

332   identified as the likely donors of three HGTs (i.e., ML00955a, ML02771a, ML18354a) in the *M.*

333   *leidyi* genome and confirmed in almost all other ctenophores transcriptomes. Other possible

334   donors could have been gymnamoebae symbionts that have been described living on the surface

335   of comb plates and on the ectoderm of ctenophores (Moss et al. 2001). However, studies

336   investigating symbiotic relationships with ctenophores are limited. Further studies are needed to

337   improve our understanding of the impacts of symbiotic relationships on HGT, as well as to

338   potentially understand the mechanisms that drive HGT between organisms.

339

340   ***Mnemiopsis leidyi* HGTs are expressed during development and encode enzymes**

341

342   Many HGTs are likely to be deleterious and lost, but some HGTs will be neutral or

343   provide a selective advantage and spread throughout a population (Thomas & Nielsen 2005).

344   HGT integration is thought to mainly occur in neutral genes with low levels of expression (Park

345   & Zhang 2012). Once integrated, neutral HGTs may become a source of novel genetic variation

346   upon which selection can act (Soucy et al. 2015). HGTs may then become more highly expressed

347   after recruitment of transcription factors and regulators from the host genome (Lercher & Pál

348   2008). Six of the nine HGTs we identified showed high expression during the first 20 hours of

349   development, suggesting potentially important developmental roles (Fig. 5). ML02771a is

350   expressed during development and encodes penicillin acylase or amidase, which catalyzes the

351   hydrolysis of benzylpenicillin. This reaction creates key intermediates for penicillin synthesis

352   and may be important to defend against microbial infection or colonization.

353

354   Observations of HGT patterns in prokaryotes have also suggested that there is a

355   preference to retain operational genes (e.g., metabolic enzymes) rather than informational genes

356   (Lawrence & Roth 1996; Jain et al. 1999; Garcia-Vallvé et al. 2000). Informational genes, such

357   as those involved in DNA replication, transcription, and translation are seldom found in sets of

358   HGTs (Thomas & Nielsen 2005). This propensity for operational genes is thought to occur

359   because informational genes are involved in larger and complex systems (Jain et al. 1999).

360   Recently, this pattern has also been observed in animal HGTs (Boto, 2014) (e.g., Zhu et al.

361   2011; Boschetti et al. 2012; Sun et al. 2013; Eyres et al. 2015; Conaco et al. 2016). These reports

362   suggest that operational genes are preferentially transferred and/or retained in both prokaryotes

363   and eukaryotes. Our data support this idea since all of the characterizable genes in our HGT set

364   encode enzymes.

365

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

367

368   Identifying HGTs can be challenging due to bacterial associations with hosts

369   (Artamonova & Mushegian 2013; Chapman et al. 2010; Fraune & Bosch 2007), DNA extraction

370   kits and reagents that have led to contamination (Naccache et al. 2013; Salter et. al 2014), and/or

371   laboratory conditions that can contaminate preparations during DNA extraction (Laurence et al.

372   2014; Strong et al. 2014). These challenges associated with sequencing and assembly have

373   resulted in contamination in public databases (Longo et al. 2011; Merchant et al. 2014) and make

374   HGT predictions difficult. Moreover, while BLAST-based approaches (i.e., alien index and the

375   HGT index) are useful for identification of HGT candidates, they are difficult to implement, lack

376   an evolutionary perspective, and do not address problems associated with contamination.

377

378   To overcome some of these challenges, we developed alien_index to automate the

379   generation of alien index and HGT index scores for rapid identification of HGT candidates. We

380   confirmed HGTs by using rigorous phylogenetic approaches to address the problems associated

381   with the lack of evolutionary perspective from BLAST methods. Our phylogenetic tests of

382   incongruence provided clear metrics from which to judge the level of certainty applied to each

383   HGT candidate. Our study showed that many of the predictions based on BLAST did not stand

384   up to hypothesis testing, and suggest that the similarity between sequences that cause high alien

385   indices do not necessarily provide true phylogenetic signal. Consequently, incorporation of

386   phylogenetic likelihood-based methods are necessary when performing HGT analyses in

387   animals.

388

389   **Conclusion**

390

391   The importance of HGT as an evolutionary mechanism in prokaryotes and eukaryotes has

392   been underestimated. While studies of HGT in animals are gradually becoming more accepted,

393   many challenges remain to quantify the extent of HGT and its impacts. To mitigate some of

394   these challenges, rigorous approaches that employ both BLAST- and phylogenetic likelihood-

395   based methods should be applied to future HGT studies in animals. Here we provided evidence

396   of nine cases of HGT in ctenophores by applying these rigorous methods (among others), and

397   found similar patterns of transfer between prokaryotes and eukaryotes with preference for

398   operational genes. It should be noted that we implemented an extremely conservative approach

399   and there are likely to be more HGTs in *M. leidyi.* However, many more studies will be

400   necessary to gain a comprehensive overview of HGT and the mechanisms by which HGT occurs

401   in animals.

402   **Acknowledgements**

403   We would like to thank Melissa DeBiasse for constructive comments on earlier versions of our

404   manuscript.

405   **References**

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

409   Anavy L., Levin M., Khair S., Nakanishi N., Fernandez-Valverde SL., Degnan BM., Yanai Y.
410   2014. BLIND ordering of large-scale transcriptomic developmental timecourses.
411   *Development* 141:1161–1166. DOI: 10.1242/DEV.105288

412   Anders S., Pyl PT., Huber W. 2015. HTSeq--a Python framework to work with high-throughput
413   sequencing data. *Bioinformatics* 31:166–169. DOI: 10.1093/bioinformatics/btu638

414   Andersson JO., Doolittle WF., Nesbø CL. 2001. Are There Bugs in Our Genome? *Science*
415   292:1848-1850.

416   Andersson JO., Hirt RP., Foster PG., Roger AJ. 2006. Evolution of four gene families with
417   patchy phylogenetic distributions: influx of genes into protist genomes. *BMC Evolutionary*
418   *Biology* 6:27. DOI: 10.1186/1471-2148-6-27

419   Arcila D., Ortí G., Vari R., Armbruster JW., Stiassny MLJ., Ko KD., Sabaj MH., Lundberg J.,
420   Revell LJ., Betancur-R. R. 2017. Genome-wide interrogation advances resolution of
421   recalcitrant groups in the tree of life. *Nature Ecology & Evolution* 1:20. DOI:
422   10.1038/s41559-016-0020

423   Artamonova II., Mushegian AR. 2013. Genome sequence analysis indicates that the model
424   eukaryote Nematostella vectensis harbors bacterial consorts. *Applied and Environmental*
425   *Microbiology* 79:6868–6873. DOI: 10.1128/AEM.01635-13

426   Artamonova II., Lappi T., Zudina L., Mushegian AR. 2015. Prokaryotic genes in eukaryotic
427   genome sequences: when to infer horizontal gene transfer and when to suspect an actual
428   microbe. *Environmental Microbiology* 17:2203–2208. DOI: 10.1111/1462-2920.12854

429   Bapteste E., Moreira D., Philippe H. 2003. Rampant horizontal gene transfer and phospho-donor
430   change in the evolution of the phosphofructokinase. *Gene* 318:185-191.

431   Bhattacharya D., Pelletreau KN., Price DC., Sarver KE. 2013. Genome Analysis of Elysia
432   chlorotica Egg DNA Provides No Evidence for Horizontal Gene Transfer into the Germ
433   Line of This Kleptoplastic Mollusc. *Molecular Biology and Evolution* 30:1843–1852. DOI:
434   10.1093/molbev/mst084

435   Boothby TC., Tenlen JR., Smith FW., Wang JR., Patanella KA., Nishimura EO., Tintori SC., Li
436   Q., Jones CD., Yandell M., Messina DN., Glasscock J., Goldstein B. 2015. Evidence for
437   extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the*

438 *National Academy of Sciences of the United States of America* 112:15976–15981. DOI:
439   10.1073/pnas.1510461112

440 Borowiec ML., Lee EK., Chiu JC., Plachetzki DC. 2015. Extracting phylogenetic signal and
441   accounting for bias in whole-genome data sets supports the Ctenophora as sister to
442   remaining Metazoa. *BMC Genomics* 16:987. DOI: 10.1186/s12864-015-2146-4

443 Boschetti C., Carr A., Crisp A., Eyres I., Wang-Koh Y., Lubzens E., Barraclough TG., Micklem
444   G., Tunnacliffe A. 2012. Biochemical Diversification through Foreign Gene Expression in
445   Bdelloid Rotifers. *PLoS Genetics* 8:e1003035. DOI: 10.1371/journal.pgen.1003035

446 Boto L. 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings.*
447   *Biological Sciences* 277:819–827. DOI: 10.1098/rspb.2009.1679

448 Boto L. 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans.
449   *Proceedings of the Royal Society of London B: Biological Sciences* 281:20132450. DOI:
450   10.1098/rspb.2013.2450

451 Buss LW. 1983. Evolution, development, and the units of selection. *Proceedings of the National*
452   *Academy of Sciences* 80:1387-1391.

453 Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in
454   phylogenetic analysis. *Molecular Biology and Evolution* 17:540-552. DOI:
455   10.1093/oxfordjournals.molbev.a026334

456 Chang ES., Neuhof M., Rubinstein ND., Diamant A., Philippe H., Huchon D., Cartwright P.
457   2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria.
458   *Proceedings of the National Academy of Sciences of the United States of America*
459   112:14912–14917. DOI:10.1073/pnas.1511468112

460 Chapman JA., Kirkness EF., Simakov O., Hampson SE., Mitros T., Weinmaier T., Rattei T.,
461   Balasubramanian PG., Borman J., Busam D., Disbennett K., Pfannkoch C., Sumin N.,
462   Sutton GG., Viswanathan LD., Walenz B., Goodstein DM., Hellsten U., Kawashima T.,
463   Prochnik SE., Putnam NH., Shu S., Blumberg B., Dana CE., Gee L., Kibler DF., Law L.,
464   Lindgens D., Martinez DE., Peng J., Wigge PA., Bertulat B., Guder C., Nakamura Y.,
465   Ozbek S., Watanabe H., Khalturin K., Hemmrich G., Franke A., Augustin R., Fraune S.,
466   Hayakawa E., Hayakawa S., Hirose M., Hwang JS., Ikeo K., Nishimiya-Fujisawa C., Ogura
467   A., Takahashi T., Steinmetz PRH., Zhang X., Aufschnaiter R., Eder M-K., Gorny A-K.,
468   Salvenmoser W., Heimberg AM., Wheeler BM., Peterson KJ., Böttger A., Tischler P., Wolf
469   A., Gojobori T., Remington KA., Strausberg RL., Venter JC., Technau U., Hobmayer B.,
470   Bosch TCG., Holstein TW., Fujisawa T., Bode HR., David CN., Rokhsar DS., Steele RE.
471   2010. The dynamic genome of Hydra. *Nature* 464:592–596. DOI: 10.1038/nature08830.

472

473   Church SH., Ryan JF., Dunn CW. 2015. Automation and Evaluation of the SOWH Test with
474   SOWHAT. *Systematic Biology* 64:1048–1058. DOI: 10.1093/sysbio/syv055

475   Conaco C., Tsoulfas P., Sakarya O., Dolan A., Werren J., Kosik K. S. 2016. Detection of
476   Prokaryotic Genes in the Amphimedon queenslandica Genome. *PLOS ONE* 11:e0151092.
477   DOI: 10.1371/journal.pone.0151092

478   Daniels C., Breitbart M. 2012. Bacterial communities associated with the ctenophores
479   Mnemiopsis leidyi and Beroe ovata. *FEMS Microbiology Ecology* 82:90–101. DOI:
480   10.1111/j.1574-6941.2012.01409.x

481   Dehal P., Satou Y., Campbell RK., Chapman J., Degnan B., De Tomaso A., Davidson B., Di
482   Gregorio A., Gelpke M., Goodstein DM., Harafuji N., Hastings KEM., Ho I., Hotta K.,
483   Huang W., Kawashima T., Lemaire P., Martinez D., Meinertzhagen IA., Necula S., Nonaka
484   M., Putnam N., Rash S., Saiga H., Satake M., Terry A., Yamada L., Wang H-G., Awazu S.,
485   Azumi K., Boore J., Branno M., Chin-bow S., DeSantis R., Doyle S., Francino P., Keys
486   DN., Haga S., Hayashi H., Hino K., Imai KS., Inaba K., Kano S., Kobayashi K., Kobayashi
487   K., Lee B-I., Makabe KW., Manohar C., Matassi G., Medina M., Mochizuki Y., Mount S.,
488   Morishita T., Miura S., Nakayama A., Nishizaka S., Nomoto H., Ohta F., Oishi K.,
489   Rigoutsos I., Sano M., Sasaki A., Sasakura Y., Shoguchi E., Shin-i T., Spagnuolo A.,
490   Stainier D., Suzuki MM., Tassy O., Takatori N., Tokuoka M., Yagi K., Yoshizaki F., Wada
491   S., Zhang C., Hyatt PD., Larimer F., Detter C., Doggett N., Glavina T., Hawkins T.,
492   Richardson P., Lucas S., Kohara Y., Levine M., Satoh N., Rokhsar DS. 2002. The Draft
493   Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science*
494   298:2157-2167. DOI: 10.1126/science.1080049

495   Delmont TO., Eren AM. 2016. Identifying contamination with advanced visualization and
496   analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*
497   4:e1839. DOI: 10.7717/peerj.1839.

498   Doolittle WF. 1999. Phylogenetic Classification and the Universal Tree. *Science* 284:2124-2128.

499   Dunn CW., Hejnol A., Matus DQ., Pang K., Browne WE., Smith SA., Seaver E., Rouse GW.,
500   Obst M., Edgecombe GD., Sørensen M V., Haddock SHD., Schmidt-Rhaesa A., Okusu A.,
501   Kristensen RM., Wheeler WC., Martindale MQ., Giribet G. 2008. Broad phylogenomic
502   sampling improves resolution of the animal tree of life. *Nature* 452:745–749. DOI:
503   10.1038/nature06614.

504   Eliáš M., Klimeš V., Derelle R., Petrželková R., Tachezy J. 2016. A paneukaryotic genomic
505   analysis of the small GTPase RABL2 underscores the significance of recurrent gene loss in
506   eukaryote evolution. *Biology direct* 11:5. DOI: 10.1186/s13062-016-0107-8

507   Eyres I., Boschetti C., Crisp A., Smith TP., Fontaneto D., Tunnacliffe A., Barraclough TG. 2015.
508   Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species
509   from desiccating habitats. *BMC Biology* 13. DOI: 10.1186/s12915-015-0202-9.

510 Fraune S., Bosch TCG. 2007. Long-term maintenance of species-specific bacterial microbiota in
511   the basal metazoan Hydra. *Proceedings of the National Academy of Sciences of the United*
512   *States of America* 104:3146–13151. DOI: 10.1073/PNAS.0703375104

513 Feuda R., Dohrmann M., Pett W., Philippe H., Rota-Stabelli O., Lartillot N., Wörheide G., Pisani
514   D. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to
515   All Other Animals. *Current Biology* 27:3864-3870. DOI: 10.1016/j.cub.2017.11.008.

516 Garcia-Vallvé S., Romeu A., Palau J. 2000. Horizontal gene transfer in bacterial and archaeal
517   complete genomes. *Genome Research* 10:1719–1725. DOI: 10.1101/GR.130000

518 Gladyshev EA., Meselson M., Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid
519   rotifers. *Science* 320:1210–1213. DOI: 10.1126/science.1156407

520 Glynn PW., Coffman B., Fuller MPC., Moorhead SG., Williams MK., Primov KD., Fortson TN.,
521   Barrales RN., Glynn PJ. 2017. Benthic ctenophores (Platyctenida: Coeloplanidae) in south
522   Florida: environmental conditions, habitats, abundances, and behaviors. *Invertebrate*
523   *Biology* 136:379-393. DOI: 10.1111/ivb.12189.

524 Haegeman A., Jones JT., Danchin EG. 2011. Horizontal gene transfer in nematodes: a catalyst
525   for plant parasitism?. *Molecular Plant-Microbe Interactions* 24:879-887.

526 Hall C., Brachat S., Dietrich, FS. 2005. Contribution of horizontal gene transfer to the evolution
527   of Saccharomyces cerevisiae. *Eukaryotic Cell* 4:1102–1115. DOI: 10.1128/EC.4.6.1102-
528   1115.2005

529 Hashimshony T., Wagner F., Sher N., Yanai I. 2012. CEL-Seq: Single-Cell RNA-Seq by
530   Multiplexed Linear Amplification. *Cell Reports* 2:666–673. DOI:
531   10.1016/J.CELREP.2012.08.003.

532 Hejnol A., Obst M., Stamatakis A., Ott M., Rouse GW., Edgecombe GD., Martinez P., Baguñà
533   J., Bailly X., Jondelius U., Wiens M., Müller WEG., Seaver E., Wheeler WC., Martindale
534   MQ., Giribet G., Dunn CW. 2009. Assessing the root of bilaterian animals with scalable
535   phylogenomic methods. *Proceedings of the Royal Society of London B: Biological Sciences*
536   276:4261–4270. DOI: 10.1098/rspb.2009.0896.

537 Jain R., Rivera MC., Lake JA. 1999. Horizontal Gene Transfer among Genomes: The
538   Complexity Hypothesis. *Proceedings of the National Academy of Sciences of the United*
539   *States of America Evolution* 96:3801–3806. DOI: 10.1073/pnas.96.7.3801

540 Jensen L., Grant JR., Laughinghouse HD., Katz LA. 2016. Assessing the effects of a sequestered
541   germline on interdomain lateral gene transfer in Metazoa. *Evolution* 70:1322–1333. DOI:
542   10.1111/evo.12935.

543 Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple
544   sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059–3066.
545   DOI: 10.1093/nar/gkf436.

546 Katoh K., Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
547   Improvements in Performance and Usability. *Molecular Biology and Evolution* 30:772–780.
548   DOI: 10.1093/molbev/mst010.

549 Katz LA. 2002. Lateral gene transfers and the evolution of eukaryotes: theories and data.
550   *International Journal of Systematic and Evolutionary Microbiology* 52:1893–1900. DOI:
551   10.1099/00207713-52-5-1893

552 Keeling PJ., Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews*
553   *Genetics* 9:605–618. DOI: 10.1038/nrg2386

554 Koutsovoulos G., Kumar S., Laetsch DR., Stevens L., Daub J., Conlon C., Maroon H., Thomas
555   F., Aboobaker AA., Blaxter M. 2016. No evidence for extensive horizontal gene transfer in
556   the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of*
557   *Sciences* 113:5053-5058. DOI: 10.1073/pnas.1600338113.

558 Lander ES., Linton LM., Birren B., Nusbaum C., Zody MC., Baldwin J., Devon K., Dewar K.,
559   Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L.,
560   Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov JP., Miranda C.,
561   Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-
562   Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough
563   R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas
564   P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T.,
565   Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne
566   S., Mullikin JC., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston RH.,
567   Wilson RK., Hillier LW., McPherson JD., Marra MA., Mardis ER., Fulton LA., Chinwalla
568   AT., Pepin KH., Gish WR., Chissoe SL., Wendl MC., Delehaunty KD., Miner TL.,
569   Delehaunty A., Kramer JB., Cook LL., Fulton RS., Johnson DL., Minx PJ., Clifton SW.,
570   Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N.,
571   Cheng J-F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs RA., Muzny
572   DM., Scherer SE., Bouck JB., Sodergren EJ., Worley KC., Rives CM., Gorrell JH., Metzker
573   ML., Naylor SL., Kucherlapati RS., Nelson DL., Weinstock GM., Sakaki Y., Fujiyama A.,
574   Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T.,
575   Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E.,
576   Robert C., Wincker P., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A.,
577   Smith DR., Doucette-Stamm L., Rubenfield M., Weinstock K., Lee HM., Dubois J., Yang
578   H., Yu J., Wang J., Huang G., Gu J., Hood L., Rowen L., Madan A., Qin S., Davis RW.,
579   Federspiel NA., Abola AP., Proctor MJ., Roe BA., Chen F., Pan H., Ramser J., Lehrach H.,
580   Reinhardt R., McCombie WR., de la Bastide M., Dedhia N., Blöcker H., Hornischer K.,
581   Nordsiek G., Agarwala R., Aravind L., Bailey JA., Bateman A., Batzoglou S., Birney E.,
582   Bork P., Brown DG., Burge CB., Cerutti L., Chen H-C., Church D., Clamp M., Copley RR.,
583   Doerks T., Eddy SR., Eichler EE., Furey TS., Galagan J., Gilbert JGR., Harmon C.,

584   Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W., Johnson LS., Jones TA.,
585   Kasif S., Kaspryzk A., Kennedy S., Kent WJ., Kitts P., Koonin E V., Korf I., Kulp D.,
586   Lancet D., Lowe TM., McLysaght A., Mikkelsen T., Moran J V., Mulder N., Pollara VJ.,
587   Ponting CP., Schuler G., Schultz J., Slater G., Smit AFA., Stupka E., Szustakowski J.,
588   Thierry-Mieg D., Thierry-Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf
589   YI., Wolfe KH., Yang S-P., Yeh R-F., Collins F., Guyer MS., Peterson J., Felsenfeld A.,
590   Wetterstrand KA., Myers RM., Schmutz J., Dickson M., Grimwood J., Cox DR., Olson M
591   V., Kaul R., Raymond C., Shimizu N., Kawasaki K., Minoshima S., Evans GA., Athanasiou
592   M., Schultz R., Patrinos A., Morgan MJ. 2001. Initial sequencing and analysis of the human
593   genome. *Nature* 409:860–921. DOI: 10.1038/35057062.

594   Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
595   9:357–359. DOI: 10.1038/nmeth.1923

596   Laurence M., Hatzis C., Brash DE. 2014. Common Contaminants in Next-Generation
597   Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLoS ONE* 9:e97876.
598   DOI: 10.1371/journal.pone.0097876

599   Lawrence JG., Roth JR. 1996. Selfish Operons: Horizontal Transfer May Drive the Evolution of
600   Gene Clusters. *Genetics* 143:1843-1860.

601   Lercher MJ., Pal C. 2008. Integration of Horizontally Transferred Genes into Regulatory
602   Interaction Networks Takes Many Million Years. *Molecular Biology and Evolution* 25:559–
603   567. DOI: 10.1093/molbev/msm283

604   Levin M., Anavy L., Cole AG., Winter E., Mostov N., Khair S., Senderovich N., Kovalev E.,
605   Silver DH., Feder M., Fernandez-Valverde SL., Nakanishi N., Simmons D., Simakov O.,
606   Larsson T., Liu S-Y., Jerafi-Vider A., Yaniv K., Ryan JF., Martindale MQ., Rink JC.,
607   Arendt D., Degnan SM., Degnan BM., Hashimshony T., Yanai I. 2016. The mid-
608   developmental transition and the evolution of animal body plans. *Nature* 531:637–641.
609   DOI: 10.1038/nature16994.

610   Longo MS., O'Neill MJ., O'Neill RJ., Clamp M., Barton G. 2011. Abundant Human DNA
611   Contamination Identified in Non-Primate Genome Databases. *PLoS ONE* 6:e16410. DOI:
612   10.1371/journal.pone.0016410.

613   Madhusoodanan J. 2015. Horizontal Gene Transfer a Hallmark of Animal Genomes? Available
614   at http://www.the-scientist.com/?articles.view/articleNo/42420/title/Horizontal-Gene-
615   Transfer-a-Hallmark-of-Animal-Genomes-/

616   Martinson EO., Martinson VG., Edwards R., Mrinalini., Werren JH. 2016. Laterally Transferred
617   Gene Recruited as a Venom in Parasitoid Wasps. *Molecular Biology and Evolution*
618   33:1042–1052. DOI: 10.1093/molbev/msv348.

619   Merchant S., Wood DE., Salzberg SL. 2014. Unexpected cross-species contamination in genome
620   sequencing projects. *PeerJ*, 2:e675. DOI: 10.7717/peerj.675

621 Moreland RT., Nguyen A-D., Ryan JF., Schnitzler CE., Koch BJ., Siewert K., Wolfsberg TG.,
622   Baxevanis AD. 2014. A customized Web portal for the genome of the ctenophore
623   Mnemiopsis leidyi. *BMC genomics* 15:316. DOI: 10.1186/1471-2164-15-316.

624 Moroz LL., Kocot KM., Citarella MR., Dosung S., Norekian TP., Povolotskaya IS., Grigorenko
625   AP., Dailey C., Berezikov E., Buckley KM., Ptitsyn A., Reshetov D., Mukherjee K., Moroz
626   TP., Bobkova Y., Yu F., Kapitonov V V., Jurka J., Bobkov Y V., Swore JJ., Girardo DO.,
627   Fodor A., Gusev F., Sanford R., Bruders R., Kittler E., Mills CE., Rast JP., Derelle R.,
628   Solovyev V V., Kondrashov FA., Swalla BJ., Sweedler J V., Rogaev EI., Halanych KM.,
629   Kohn AB. 2014. The ctenophore genome and the evolutionary origins of neural systems.
630   *Nature* 510:109–114. DOI: 10.1038/nature13400.

631 Moss AG., Estes AM., Muellner LA., Morgan DD. 2001. Protistan epibionts of the ctenophore
632   *Mnemiopsis mccradyi* Mayer. *Hydrobiologia* 451:295–304.

633 Naccache SN., Greninger AL., Lee D., Coffey LL., Phan T., Rein-Weston A., Aronsohn A.,
634   Hackett J., Delwart EL., Chiu CY. 2013. The perils of pathogen discovery: origin of a novel
635   parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of*
636   *virology* 87:11966–11977. DOI: 10.1128/JVI.02323-13.

637 O'Leary NA., Wright MW., Brister JR., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse
638   B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinkova O., Brover
639   V., Chetvernin V., Choi J., Cox E., Ermolaeva O., Farrell CM., Goldfarb T., Gupta T., Haft
640   D., Hatcher E., Hlavina W., Joardar VS., Kodali VK., Li W., Maglott D., Masterson P.,
641   McGarvey KM., Murphy MR., O'Neill K., Pujar S., Rangwala SH., Rausch D., Riddick
642   LD., Schoch C., Shkeda A., Storz SS., Sun H., Thibaud-Nissen F., Tolstoy I., Tully RE.,
643   Vatsan AR., Wallin C., Webb D., Wu W., Landrum MJ., Kimchi A., Tatusova T., DiCuccio
644   M., Kitts P., Murphy TD., Pruitt KD. 2016. Reference sequence (RefSeq) database at
645   NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids*
646   *Research* 44:D733–D745. DOI: 10.1093/nar/gkv1189.

647 Park C., Zhang J. 2012. High Expression Hampers Horizontal Gene Transfer. *Genome Biology*
648   *and Evolution* 4:523–532. DOI: 10.1093/gbe/evs030

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

651 Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard
652   E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson
653   DJ., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009.
654   Phylogenomics revives traditional views on deep animal relationships. *Current biology*
655   19:706–12. DOI: 10.1016/j.cub.2009.02.052.

656 Pible O., Hartmann EM., Imbert G., Armengaud J. 2014. The importance of recognizing and
657   reporting sequence database contamination for proteomics. *EuPA Open Proteomics* 3:246–
658   249. DOI: 10.1016/J.EUPROT.2014.04.001.

659 Pick KS., Philippe H., Schreiber F., Erpenbeck D., Jackson DJ., Wrede P., Wiens M., Alié A.,
660   Morgenstern B., Manuel M., Wörheide G. 2010. Improved phylogenomic taxon sampling
661   noticeably affects nonbilaterian relationships. *Molecular biology and evolution* 27:1983–7.
662   DOI: 10.1093/molbev/msq089.

663 Pisani D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N., Wörheide
664   G. 2015. Genomic data do not support comb jellies as the sister group to all other animals.
665   *Proceedings of the National Academy of Sciences of the United States of America*
666   112:15402–7. DOI: 10.1073/pnas.1518127112.

667 Pundir S., Martin MJ., O'Donovan C. 2017. Uniprot protein knowledgebase. *Protein*
668   *Bioinformatics: From Protein Modifications and Networks to Proteomics*, 41-55.

669 Radzvilavicius AL., Hadjivasiliou Z., Pomiankowski A., Lane N. 2016. Selection for
670   Mitochondrial Quality Drives Evolution of the Germline. *PLoS biology* 14:e2000410. DOI:
671   10.1371/journal.pbio.2000410.

672 Richards TA., Dacks JB., Campbell SA., Blanchard JL., Foster PG., McLeod R., Roberts CW.
673   2006. Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal
674   gene transfer, and endosymbiotic replacements. *Eukaryotic cell* 5:1517–31. DOI:
675   10.1128/EC.00106-06.

676 Ryan JF., Pang K., Schnitzler CE., Nguyen A-D., Moreland RT., Simmons DK., Koch BJ.,
677   Francis WR., Havlak P., Smith SA., Putnam NH., Haddock SHD., Dunn CW., Wolfsberg
678   TG., Mullikin JC., Martindale MQ., Baxevanis AD. 2013. The Genome of the Ctenophore
679   *Mnemiopsis leidyi* and Its Implications for Cell Type Evolution. *Science* 342:1242592. DOI:
680   10.1126/science.1242592

681 Salter SJ., Cox MJ., Turek EM., Calus ST., Cookson WO., Moffatt MF., Turner P., Parkhill J.,
682   Loman NJ., Walker AW. 2014. Reagent and laboratory contamination can critically impact
683   sequence-based microbiome analyses. *BMC Biology* 12:87. DOI: 10.1186/s12915-014-
684   0087-z.

685 Salzberg SL., White O., Peterson J., Eisen JA. 2001. Microbial Genes in the Human Genome:
686   Lateral Transfer or Gene Loss? *Science* 292:1903-1906. DOI: 10.1126/science.1061036

687 Schönknecht G., Weber APM., Lercher MJ. 2014. Horizontal gene acquisitions by eukaryotes as
688   drivers of adaptive evolution. *BioEssays*, 36:9–20. DOI: 10.1002/bies.201300095

689 Shen X-X., Hittinger CT., Rokas A. 2017. Contentious relationships in phylogenomic studies can
690   be driven by a handful of genes. *Nature ecology & evolution* 1:0126. DOI: 10.1038/s41559-
691   017-0126

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

694 Simion P., Philippe H., Baurain D., Jager M., Richter DJ., Di Franco A., Roure B., Satoh N.,
695   Quéinnec É., Ereskovsky A., Lapébie P., Corre E., Delsuc F., King N., Wörheide G.,
696   Manuel M. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the
697   Sister Group to All Other Animals. *Current Biology* 27:958–967. DOI:
698   10.1016/j.cub.2017.02.031.

699 Song J-I., Hwang S-J. 2010. New Species of Genus *Coeloplana* (Ctenophora: Tentaculata:
700   Platyctenida) from Korea. *Animal Systematics, Evolution and Diversity* 26:217–221. DOI:
701   10.5635/KJSZ.2010.26.3.217

702 Soucy SM., Huang J., Gogarten JP. 2015. Horizontal gene transfer: building the web of life.
703   *Nature Reviews Genetics* 16:472–482. DOI: 10.1038/nrg3962

704 Stamatakis A. 2014 RAxML Version 8: A tool for phylogenetic analysis and post-analysis of
705   large phylogenies. *Bioinformatics* 30:1312–1313.

706 Stanhope MJ., Lupas A., Italia MJ., Koretke KK., Volker C., Brown JR. 2001. Phylogenetic
707   analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*
708   411:940–944. DOI: 10.1038/35082058.

709 Strong MJ., Xu G., Morici L., Bon-Durant SS., Baddoo M., Lin Z., Fewell C., Taylor CM.,
710   Flemington EK. 2014. Microbial Contamination in Next Generation Sequencing:
711   Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens*
712   10:e1004437. DOI: 10.1371/journal.ppat.1004437.

713 Sun BF., Xiao JH., He SM., Liu L., Murphy RW., Huang DW. 2013. Multiple ancient horizontal
714   gene transfers and duplications in lepidopteran species. *Insect Molecular Biology* 22:72–87.
715   DOI: 10.1111/imb.12004.

716 Telford MJ., Budd GE., Philippe H. 2015. Phylogenomic Insights into Animal Evolution.
717   *Current Biology* 25:R876–R887. DOI: 10.1016/j.cub.2015.07.060

718 Thomas CM., Nielsen KM. 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer
719   between Bacteria. *Nature Reviews Microbiology* 3:711–721. DOI: 10.1038/nrmicro1234

720 Torruella G., de Mendoza A., Grau-Bové X., Antó M., Chaplin MA., del Campo J., Eme L.,
721   Pérez-Cordón G., Whipps CM., Nichols KM., Paley R., Roger AJ., Sitjà-Bobadilla A.,
722   Donachie S., Ruiz-Trillo I. 2015. Phylogenomics Reveals Convergent Evolution of
723   Lifestyles in Close Relatives of Animals and Fungi. *Current Biology* 25:2404–2410. DOI:
724   10.1016/j.cub.2015.07.053.

725 Whelan NV., Kocot KM., Moroz LL., Halanych KM. 2015. Error, signal, and the placement of
726   Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences of
727   the United States of America* 112:5773–8. DOI: 10.1073/pnas.1503453112.

728   Whelan NV., Kocot KM., Moroz TP., Mukherjee K., Williams P., Paulay G., Moroz LL.,
729   Halanych KM. 2017. Ctenophore relationships and their placement as the sister group to all
730   other animals. *Nature Ecology & Evolution* 1:1737–1746. DOI: 10.1038/s41559-017-0331-
731   3.

732   Wybouw N., Pauchet Y., Heckel DG., Van Leeuwen T. 2016. Horizontal Gene Transfer
733   Contributes to the Evolution of Arthropod Herbivory. *Genome biology and evolution*
734   8:1785–801. DOI: 10.1093/gbe/evw119.

735   Zhu B., Lou M-M., Xie G-L., Zhang G-Q., Zhou X-P., Li B., Jin G-L. 2011. Horizontal gene
736   transfer in silkworm, *Bombyx mori*. *BMC Genomics* 12:248. DOI: 10.1186/1471-2164-12-
737   248.

Pipeline and outputs to identify and confirm HGTs.

(A) alien_index was used to identify 37 HGT candidates. (B) These candidates were then BLASTed against RefSeq; two candidates were removed because they only had two significant animals hits and six were set aside for future testing because they lacked animal hits. (C) The remaining 29 candidates were tested by phylogenetic analyses and hypothesis testing (SOWH and AU test). The 6 candidates that lacked animal hits and those that passed hypothesis testing (7 candidates) were screened for significant hits to Choanoflagellates.

A. alien_index

ML2.2
16,548 gene models

BLASTP

alien_index database

Animals

non-animal Eukaryotes

Bacteria

Archaea

alien_index

37 candidates

B. RefSeq

BLASTP

RefSeq

Animals

non-animal Eukaryotes

Fungi

Bacteria

C. Hypothesis testing

phylogenetic criteria

RAxML

MAFFT/ GBlocks

12 candidates

29 gene trees

29 alignments

set of related seqs for 29 candidates

two top animal hits

2 removed

SOWH/ AU-test

7 candidates

Choano present?

6 candidates

3 candidates

Choano present?

6 Seqs w/ no animal hits

9 Confirmed HGTs

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 (E) are examples of RAxML best trees for HGT candidates validated 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. The asterisk indicates a gene that is later removed from contention.

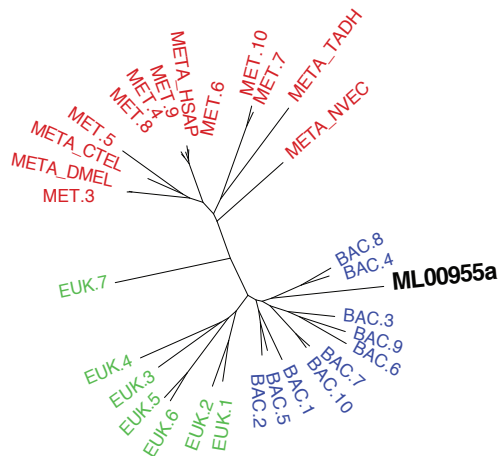
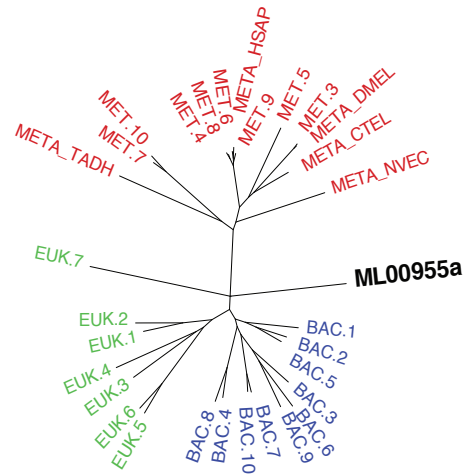
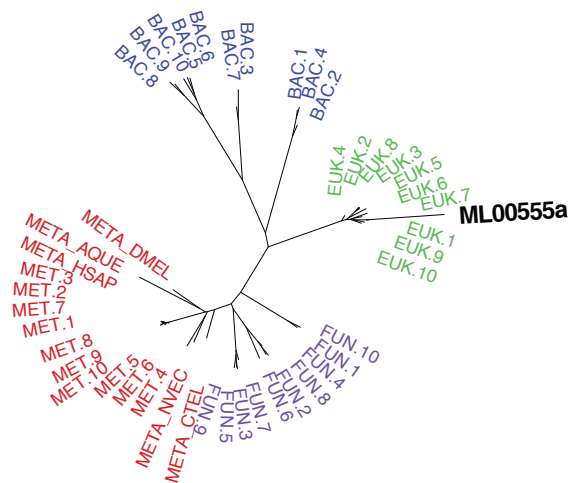
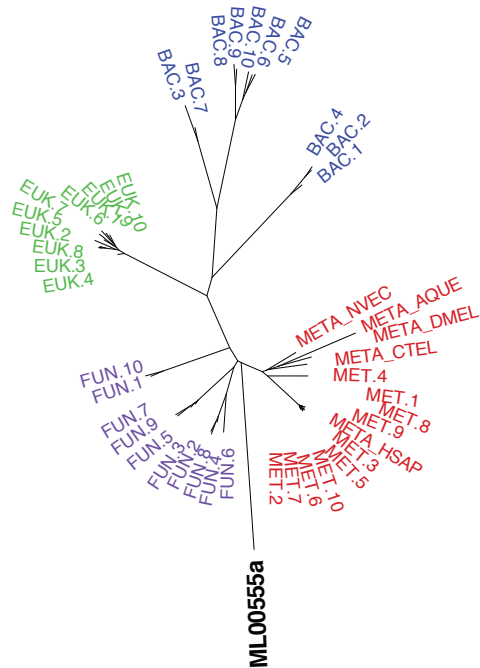
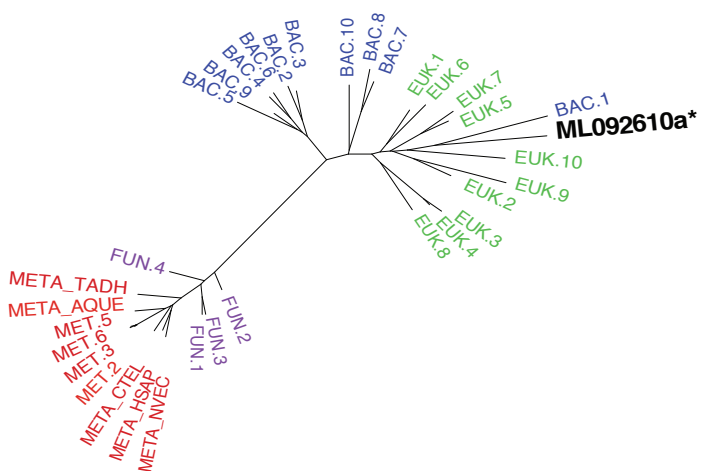
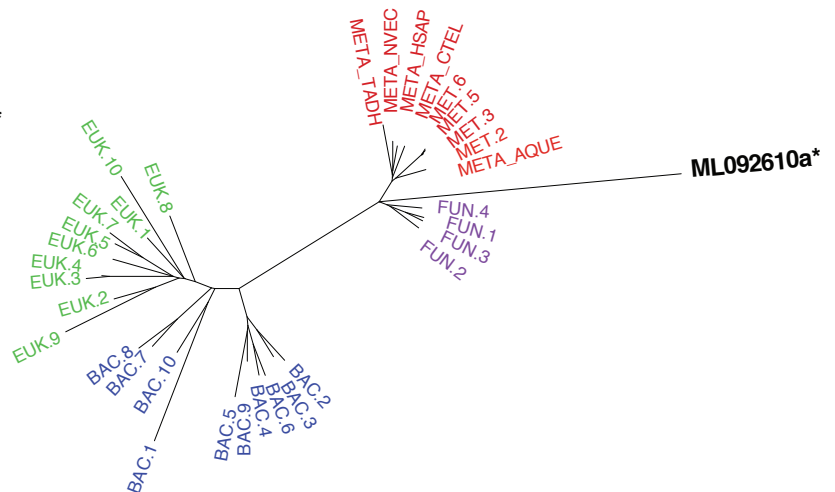
A

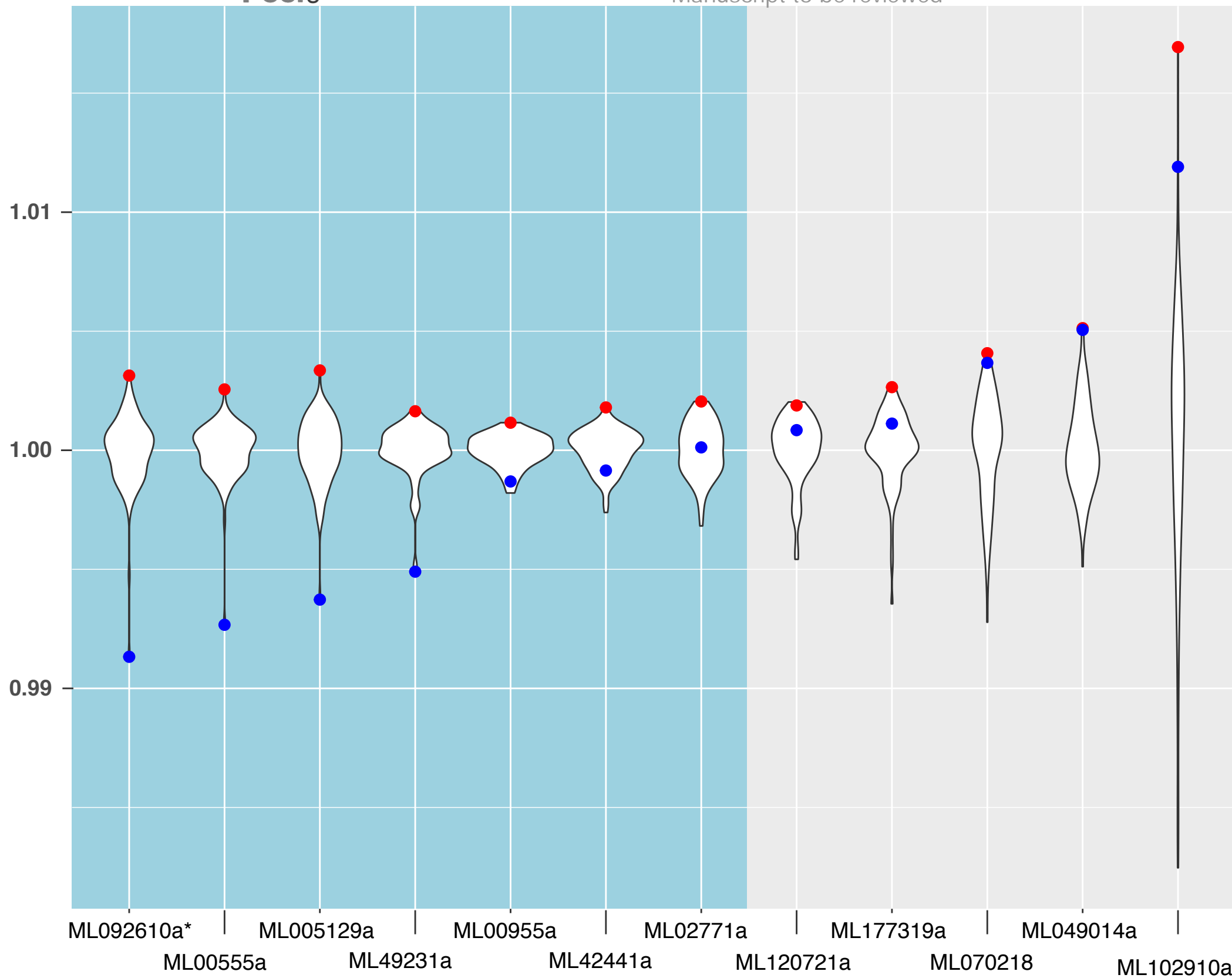
B

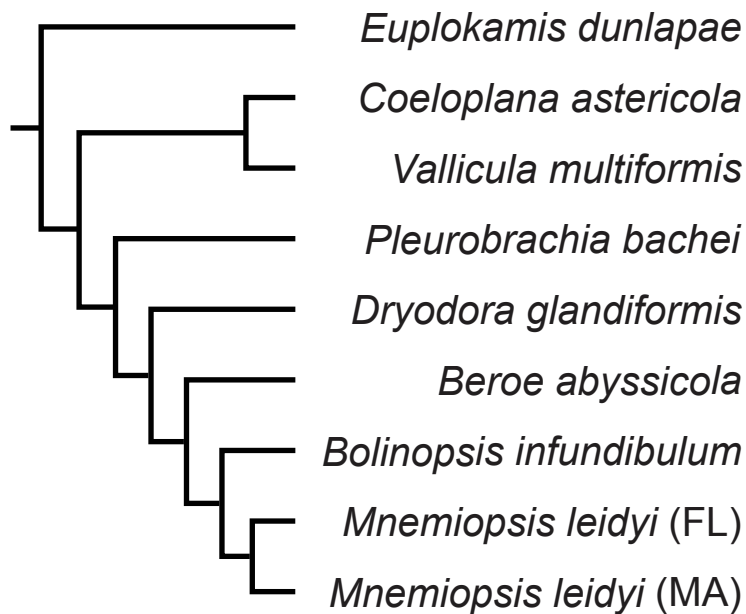
C

D

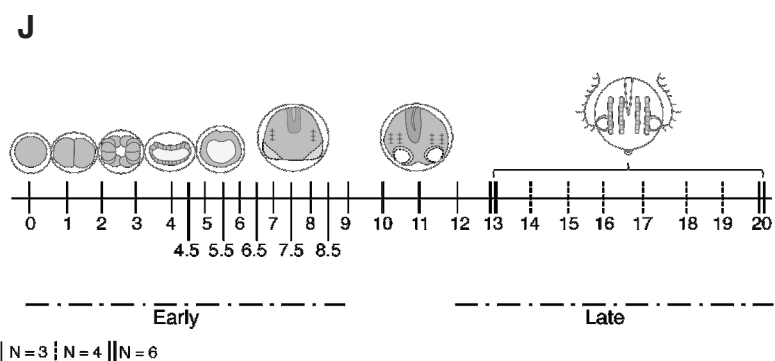
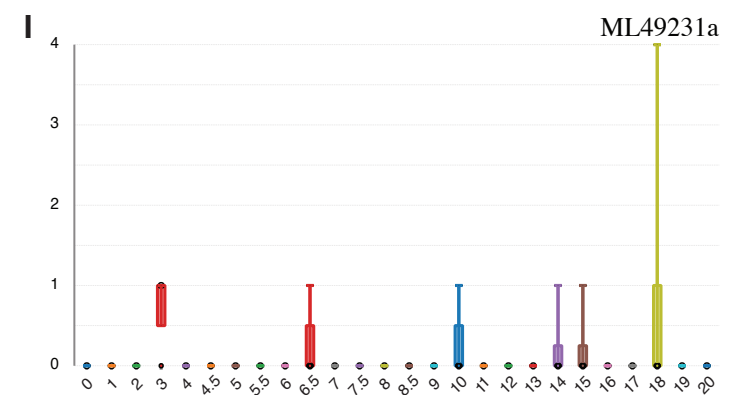
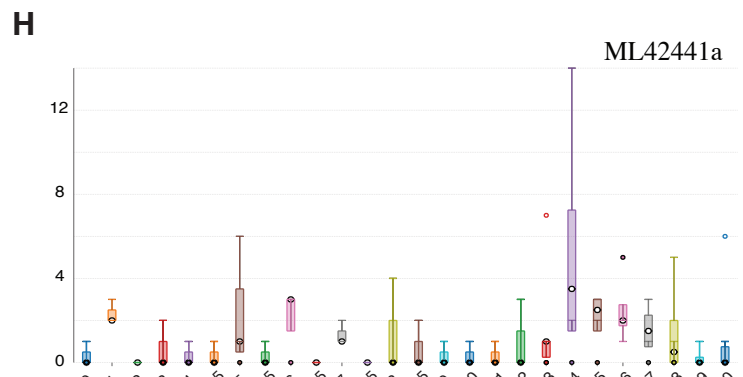
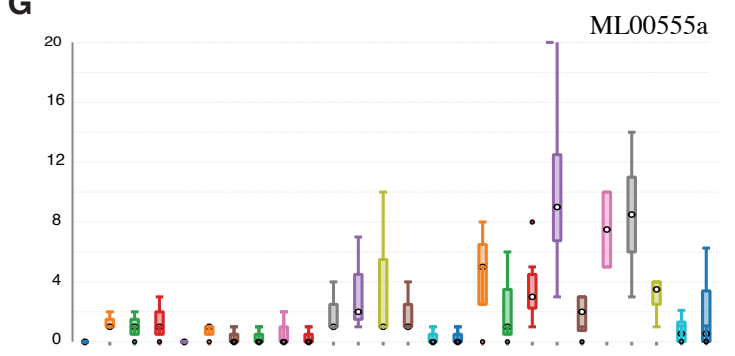
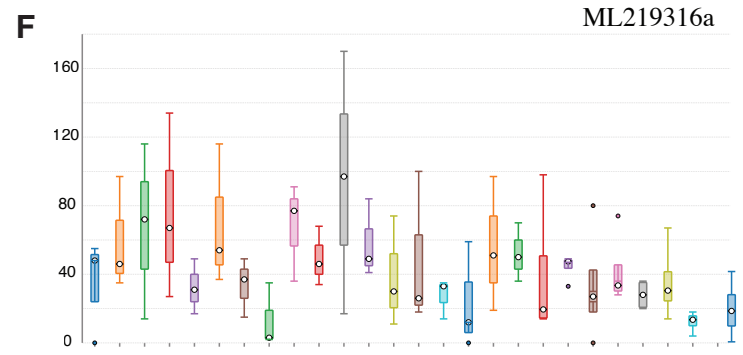
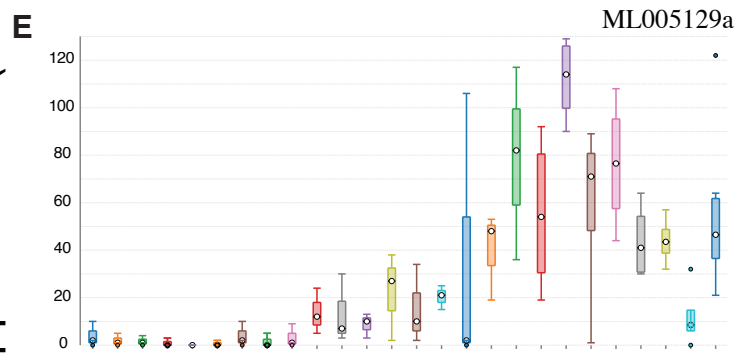
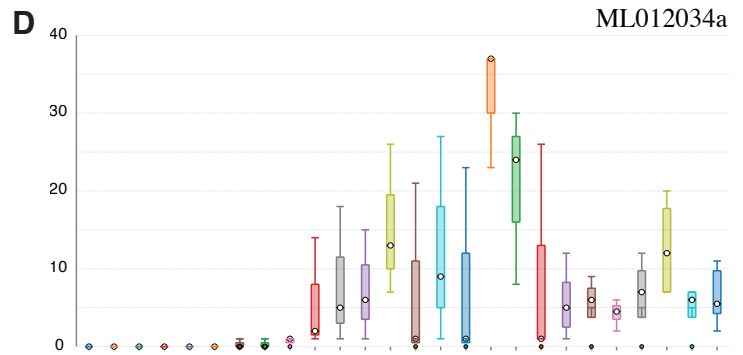
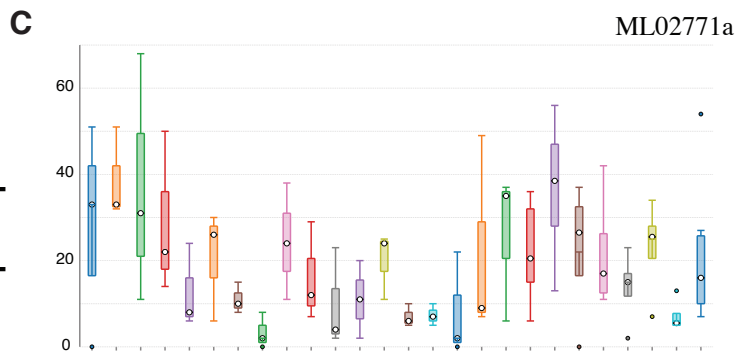
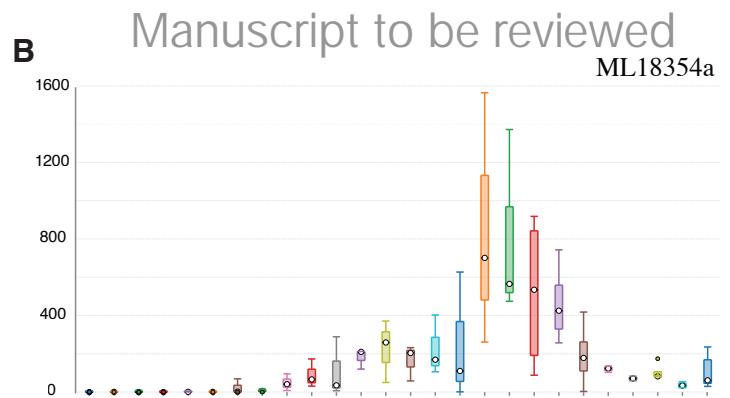
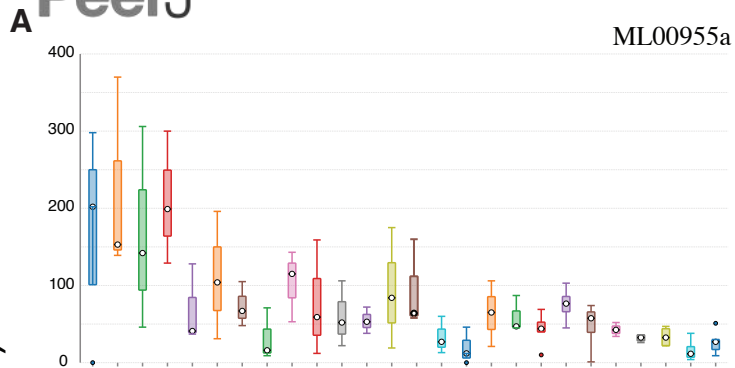
E

F

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. The side in teal shows HGT candidates validated by phylogenetic analyses and hypothesis testing. The asterisk indicates a gene that is later removed from contention.

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). Percent identity among genes are described in Table S4.

Expression profiles of the nine HGTs identified in this study.

Single-embryo RNA-Seq analyses were performed over 20 hours. (A-F) 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. (G-I) Confirmed HGTs with transcript-per-million values less than 100 over 20 hours. (J) Ctenophore stages of development over the timecourse. Early cleavage stages occur at 1-3 hpf. Gastrulation occurs at 4-6 hpf . Tentacle morphogenesis occurs at 9-12 hpf. N refers to the number of replicates.

Number of Mapped Reads (Transcripts-per-million)

Hours post fertilization

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) likely HGTs. The asterisk indicates a gene that is later removed from contention.

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

Intron structure of nine HGTs and surrounding genes.

The genes highlighted in red are the HGT candidates. The gene with an asterisk indicates one non-spliceosomal intron.

Candidate HGT	Number of introns in candidate HGTs and surrounding genes						
ML00955a	ML00952a 5	ML00953a 0	ML00954a 4	ML00955a 1	ML00956a 0	ML00957a 7	ML00958a 2
ML18354a	ML18351a 6	ML18352a 14	ML18353a 7	ML18354a 5	ML18355a 1	ML18356a 0	ML18357a 0
ML02771a				ML02771a 7	ML02772a 8	ML02773a 16	ML02774a 3
ML012034a	ML012031a 6	ML012032a 5	ML012033a 6	ML012034a 0	ML012035a 3	ML012036a 6	
ML005129a	ML005126a 0	ML005127a 7	ML005128a 2	ML005129a 1	ML005130a 0	ML005131a 4	ML005132a 8
ML219316a	ML219313a 6	ML219314a 3	ML219315a 7*	ML219316a 4	ML219317a 6		
ML00555a	ML00552a 0	ML00553a 12	ML00554a 0	ML00555a 14	ML00556a 3	ML00557a 0	ML00558a 5
ML42441a				ML42441a 1	ML42442a 14	ML42443a 3	ML42444a 10
ML49231a				ML49231a 6			

Sumary of confirmed 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 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