

Supplementary text: commands and parameters for phylogenetic analyses

All files and scripts are available at:

https://github.com/josephryan/2017-Hernandez_and_Ryan

1. Alignments

Alignments were generated using MAFFT version 7.305 and trimmed with Gblockswrapper version 0.91b using the following commands:

```
mafft seqs.fa > aln.fa
```

```
Gblockswrapper aln.fa
```

Gblockswrapper is available here:

https://bitbucket.org/caseywdunn/labcode/src/master/scripts_phylogenomics_21Feb2009/Gblockswrapper

```
perl -pi -e 's/ //g' aln.fa-gb
```

```
fasta2phy.pl aln.fa-gb > aln.fa-gb.phy
```

fasta2phy.pl is a script bundled with JFR-PerlModules (Release 1.0+)

<https://github.com/josephryan/JFR-PerlModules/>

2. Tree inference

We performed maximum likelihood analyses on the alignments using RAxML version 8.1.21:

```
raxmlHPC -p 1234 -m PROTGAMMAGTR -n ALN -s aln.fa-gb.phy
```

3. Hypothesis testing

All commands used for hypothesis testing are in the script `test_hypothesized_phylogenies` uploaded to GitHub . We used the SOWH test as implemented in SOWHAT version 0.36 and the AU test implemented in CONSEL version 0.20. The script `test_hypothesized_phylogenies` generates the metazoan constraint tree as well as bootstrap trees to test as alternative hypotheses against the best tree (i.e., tree indicating HGT) produced by RAxML. The following command is used run the script:

```
perl test_hypothesized_phylogenies.pl aln.fa-gb.phy dir/
```

4. Single copies of bootstrap trees

To make sure that each bootstrap tree only appeared once in a set of 100 generated bootstrap trees (i.e., suboptimal trees) we used the ape package in R with the following commands:

```
library(ape)
```

```
bs<-read.tree(file='RAxML_bootstrap.ALN.bs')
```

```
unique.multiPhylo(bs)
```

5. Manually generated suboptimal trees

We used suboptimal trees to address the problem of selection bias which occurs when using the AU test. We manually created suboptimal trees in order to get a feel for how different bootstrap trees were from the optimal tree. Suboptimal trees were manually created by switching the positions of taxa in clades consisting of three taxa from the best tree (Fig. S1). We included 10 manually generated suboptimal trees, or in cases with less than 10 trees, as many as could be generated in each analysis for the AU test. To optimize manually generated suboptimal trees we performed maximum likelihood analyses before implementing the AU test. All manually created suboptimal trees have been uploaded to the accompanying GitHub site.

We used the following commands to optimize each manually generated suboptimal tree:

```
raxmlHPC -p 1234 -m PROTGAMMAGTR -n subtree1 -s aln.fa-gb.phy -g  
RAxMLbestTree1_equalbranch.ALN
```

To run the AU test implemented through CONSEL on the best, metazoan constraint, and manually generated suboptimal trees:

```
cat RAxML_bestTree.ALN RAxML_bestTree.metatree RAxML_bestTree.suboptree1  
RAxML_bestTree.suboptree2 RAxML_bestTree.suboptree3  
RAxML_bestTree.suboptree4 RAxML_bestTree.suboptree5  
RAxML_bestTree.suboptree6 RAxML_bestTree.suboptree7  
RAxML_bestTree.suboptree8 RAxML_bestTree.suboptree9  
RAxML_bestTree.suboptree10 >> 12trees.tre
```

```
raxmlHPC -f g -m PROTGAMMAGTR -n 12trees -s aln.fa-gb.phy -z 12trees.tre
```

```
seqmt --puzzle RAxML_perSiteLLs.12trees
```

```
makermt RAxML_perSiteLLs
```

consel RAxML_perSiteLLs

catpv RAxML_perSiteLLs

6. Violin plots

We used violin plots to visualize comparisons of likelihood scores between the best tree (i.e., tree indicating HGT), metazoan constraint tree, and suboptimal trees. Violin plots were generated using the script `likelihood_violins` available on the accompanying GitHub site. We also made comparisons between the tree space covered by using bootstrap trees as suboptimal trees (Fig. 3) versus manually created suboptimal trees in the AU test (Fig. S3). We found that the bootstrap trees covered a wider range of tree space than the manually created trees. Therefore, bootstrap trees provide a more stringent test.

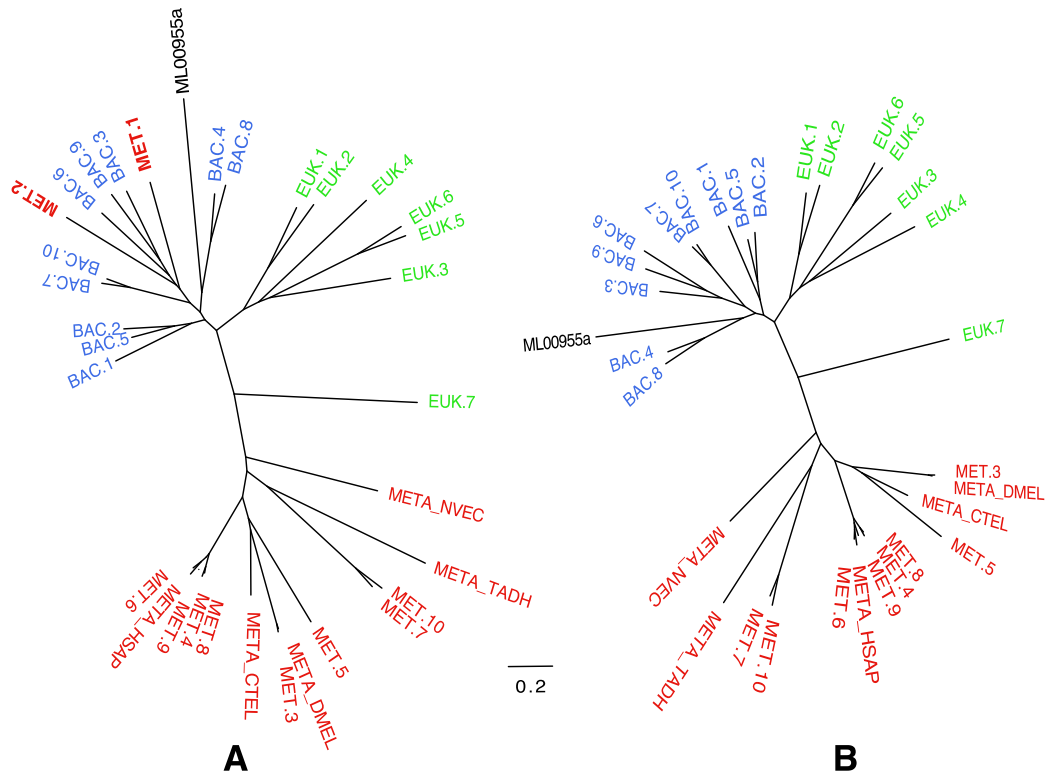


Figure S1. Maximum-likelihood analyses on an HGT candidate that includes metazoan sequences outside of the main metazoan clade. ML00955a (in black) is the *M. leidyi* HGT candidate. (A) Because there were only two non-*Mnemiopsis* animal sequences outside of the main animal clade (i.e. MET.1 and MET.2), these sequences were considered potential contaminants and removed. (B) RAxML analysis on the same alignment after pruning MET.1 (*Pantholops hodgsonii*) and MET.2 (*Caenorhabditis remanei*). 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*). MET=Metazoa; BAC=Bacteria; EUK=Eukaryota; FUN=Fungi; More details for each taxa are specified in Table S3.

Species	Lineage
<i>Acanthamoeba castellanii</i>	Amoebozoa
<i>Aspergillus nidulans</i>	Fungi
<i>Candidatus aquiluna sp imcc13023</i>	Bacteria / Terrabacteria
<i>Candidatus nitrosopumilus salaria bd31</i>	Archaea
<i>Candidatus pelagibacter sp imcc9063</i>	Bacteria / Proteobacteria
<i>Glaciecola pallidula dsm 14239 acam 615</i>	Bacteria / Proteobacteria
<i>Marine gamma proteobacterium htcc2080</i>	Bacteria / Proteobacteria
<i>Marine group i thaumarchaeote scgc aaa799 b03</i>	Archaea
<i>Marinobacter adhaerens hp15</i>	Bacteria / Proteobacteria
<i>Phaeodactylum tricornutum</i>	Stramenopiles
<i>Saccharomyces cerevisiae</i>	Fungi
<i>Thalassiosira pseudonana</i>	Stramenopiles
<i>Amphimedon queenslandica</i>	Animal / Porifera
<i>Capitella teleta</i>	Animal / Annelida
<i>Crassostrea gigas</i>	Animal / Mollusca
<i>Daphnia pulex</i>	Animal / Arthropoda
<i>Drosophila melanogaster</i>	Animal / Arthropoda
<i>Helobdella robusta</i>	Animal / Annelida
<i>Lottia gigantea</i>	Animal / Mollusca
<i>Nematostella vectensis</i>	Animal / Cnidaria
<i>Strigamia maritima</i>	Animal / Arthropoda
<i>Strongylocentrotus purpuratus</i>	Animal / Echinodermata
<i>Trichoplax adhaerens</i>	Animal / Placozoa
<i>Homo sapiens</i>	Animal/ Chordata

Table S1. Species used in initial alien_index run to identify HGT candidates. This database is available here: http://ryanlab.whitney.ufl.edu/downloads/alien_index/

Genes	Introns	Scaffolds with intron-containing genes
ML012034a		<input type="checkbox"/>
ML03277a	<input type="checkbox"/>	<input type="checkbox"/>
ML23958a	<input type="checkbox"/>	<input type="checkbox"/>
ML005129a	<input type="checkbox"/>	<input type="checkbox"/>
ML18354a	<input type="checkbox"/>	<input type="checkbox"/>
ML02232a	<input type="checkbox"/>	<input type="checkbox"/>
ML00955a	<input type="checkbox"/>	<input type="checkbox"/>
ML42441a	<input type="checkbox"/>	<input type="checkbox"/>
ML06718a	<input type="checkbox"/>	<input type="checkbox"/>
ML092610a	<input type="checkbox"/>	<input type="checkbox"/>
ML02771a	<input type="checkbox"/>	<input type="checkbox"/>
ML49231a	<input type="checkbox"/>	
ML00555a	<input type="checkbox"/>	<input type="checkbox"/>
ML120721a	<input type="checkbox"/>	<input type="checkbox"/>
ML073257a	<input type="checkbox"/>	<input type="checkbox"/>
ML227811a		<input type="checkbox"/>
ML207910a	<input type="checkbox"/>	<input type="checkbox"/>
ML03547a	<input type="checkbox"/>	<input type="checkbox"/>
ML102221a		<input type="checkbox"/>
ML00881a		<input type="checkbox"/>
ML00882a		<input type="checkbox"/>
ML219316a	<input type="checkbox"/>	<input type="checkbox"/>
ML049014a	<input type="checkbox"/>	<input type="checkbox"/>
ML296211a		<input type="checkbox"/>
ML1541114a	<input type="checkbox"/>	<input type="checkbox"/>
ML085726a	<input type="checkbox"/>	<input type="checkbox"/>
ML009115a		<input type="checkbox"/>
ML21002a	<input type="checkbox"/>	<input type="checkbox"/>
ML132017a	<input type="checkbox"/>	<input type="checkbox"/>
ML177319a		<input type="checkbox"/>
ML019144a		<input type="checkbox"/>
ML102910a	<input type="checkbox"/>	<input type="checkbox"/>
ML018031a	<input type="checkbox"/>	<input type="checkbox"/>
ML046416a	<input type="checkbox"/>	<input type="checkbox"/>
ML22167a	<input type="checkbox"/>	<input type="checkbox"/>
ML00556a	<input type="checkbox"/>	<input type="checkbox"/>
ML070218a		<input type="checkbox"/>

Table S2. Intron/Exon Structure of HGTs. Check marks indicate the presence of introns for HGT candidates and/or candidates found on scaffolds with other intron-containing genes. Genes in blue are HGTs confirmed by phylogenetic analyses and hypothesis testing.

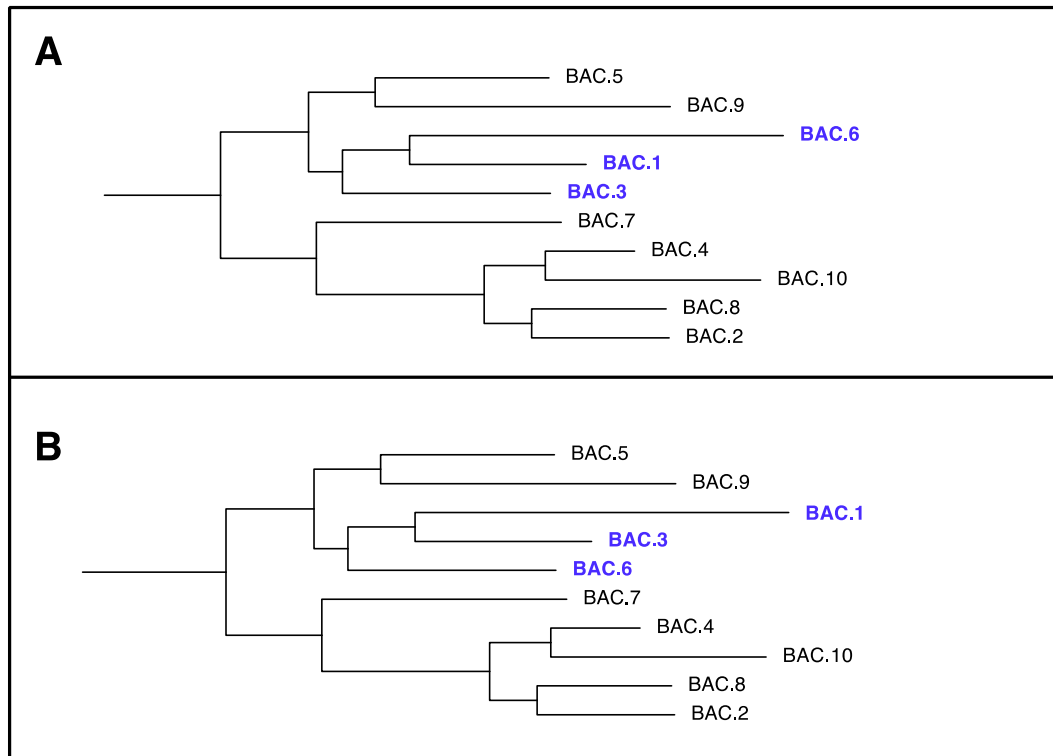


Figure S2. Examples of manually generated suboptimal trees. Suboptimal trees were manually generated by taking the best RAxML tree from each alignment and switching the positions of taxa in clades of three. All branch lengths were made equal when performing taxa switches and then optimized using RAxML. Blue text in the figure highlights the clade of three taxa in which positions of taxa will be rearranged. (A) Clade of bacteria resulting from the best maximum-likelihood tree. (B) Manually generated suboptimal tree that resulted from rearrangement of taxa highlighted in blue from the clade in (A).

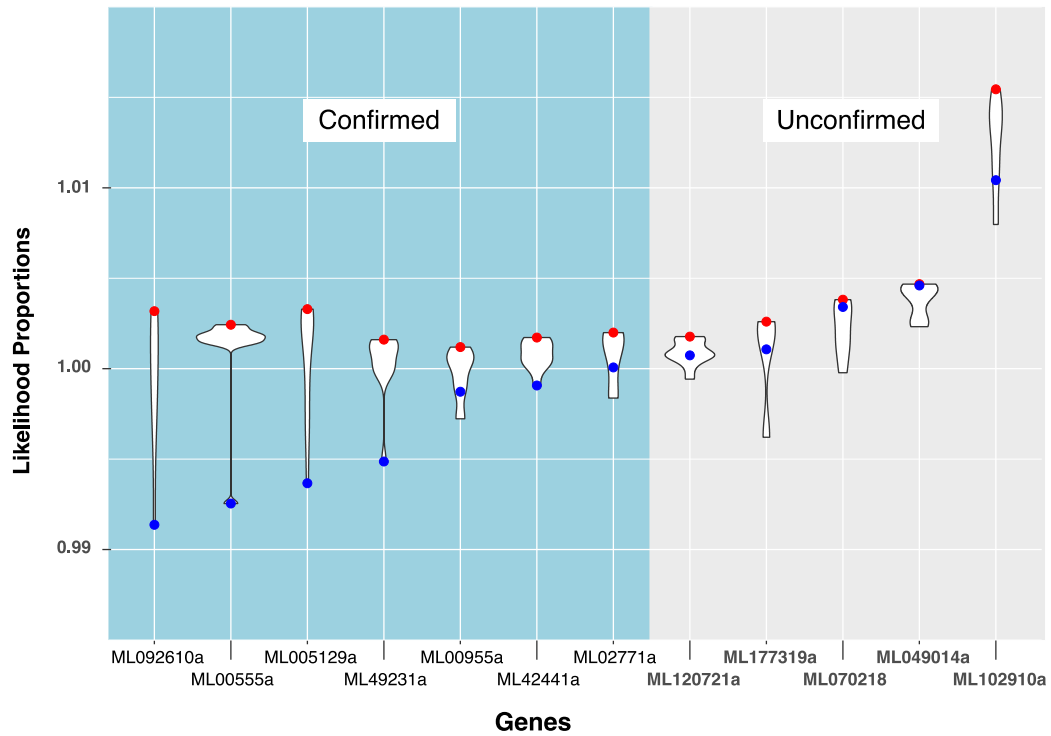


Figure S3. Manually generated suboptimal trees. Suboptimal distributions are a result of manually shuffling clades of three taxa. Results from analyses using bootstrap trees (Fig. 3) and manually created suboptimal trees are similar, but the spread of distribution of bootstrap trees is greater, making conclusions derived from comparisons to the bootstrap distributions more conservative. 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). Confirmed HGTs were validated by phylogenetic analyses and hypothesis testing.

Label	Species	Accession no.
Bac. 1	<i>Solitalea canadensis</i>	WP_014679673.1
Bac. 2	<i>Pontibacter actiniarum</i>	WP_025607756.1
Bac. 3	<i>Microbulbifer agarilyticus</i>	WP_010132679.1
Bac. 4	<i>Stenotrophomonas</i>	WP_055768138.1
Bac. 5	<i>Rufibacter</i> sp. DG31D	WP_053093794.1
Bac. 6	<i>Idiomarina zobellii</i>	WP_053954579.1
Bac. 7	<i>Echinicola vietnamensis</i>	WP_015263982.1
Bac. 8	<i>Arenimonas metalli</i>	WP_052575499.1
Bac. 9	<i>Arsukibacterium</i> sp. MJ3	WP_046552654.1
Bac. 10	<i>Anditalea andensis</i>	WP_035071114.1
Euk. 1	<i>Thalassiosira pseudonana</i> CCMP1335	XP_002296777.1
Euk. 2	<i>Guillardia theta</i> CCMP2712	XP_005831049.1
Euk. 3	<i>Micromonas</i> sp. RCC299	XP_002501910.1
Euk. 4	<i>Chondrus crispus</i>	XP_005713402.1
Euk. 5	<i>Saprolegnia parasitica</i> CBS 223.65	XP_012200393.1
Euk. 6	<i>Aphanomyces invadans</i>	XP_008879597.1
Euk. 7	<i>Galdieria sulphuraria</i>	XP_005705561.1
Met. 3	<i>Drosophila willistoni</i>	XP_015033401.1
Met. 4	<i>Protobothrops mucrosquamatus</i>	XP_015682900.1
Met. 5	<i>Helobdella robusta</i>	XP_009014349.1
Met. 6	<i>Serinus canaria</i>	XP_009093895.1
Met. 7	<i>Apis florea</i>	XP_012343105.1
Met. 8	<i>Python bivittatus</i>	XP_007437127.1, XP_007437128.1, XP_007437129.1
Met. 9	<i>Gekko japonicus</i>	XP_015268057.1
Met. 10	<i>Halyomorpha halys</i>	XP_014272262.1

Table S3A. Taxa details from Figure 2(A), (B), and Figure S1. Labels indicate the taxa labels in Figures 2(A), 2(B), and S1. Each label specifies the species name and accession number downloaded from RefSeq.

Label	Species	Accession no.
Bac. 1	<i>Bordetella</i> sp. N	WP_057653136.1
Bac. 2	<i>Lysobacter</i> sp. Root690	WP_056114344.1
Bac. 3	<i>Amycolatopsis methanolica</i>	WP_017981605.1
Bac. 4	<i>Achromobacter xylosoxidans</i>	WP_013396304.1
Bac. 5	<i>Knoellia flava</i>	WP_052116956.1
Bac. 6	<i>Nocardioides</i> sp. Soil774	WP_056601935.1
Bac. 7	<i>Actinosynnema mirum</i>	WP_012783612.1
Bac. 8	<i>Arthrobacter</i> sp. MA-N2	WP_028266024.1
Bac. 9	<i>Curtobacterium</i> sp. Leaf261	WP_055954596.1
Bac. 10	<i>Phycococcus</i> sp. Soil803	WP_057377348.1
Euk. 1	<i>Cucumis melo</i>	XP_008448951.1
Euk. 2	<i>Medicago truncatula</i>	XP_003611170.1
Euk. 3	<i>Glycine max</i>	XP_003517450.1
Euk. 4	<i>Tarenaya hassleriana</i>	XP_010540447.1
Euk. 5	<i>Beta vulgaris</i> subsp. <i>vulgaris</i>	XP_010678653.1
Euk. 6	<i>Musa acuminata</i> subsp. <i>malaccensis</i>	XP_009398075.1
Euk. 7	<i>Cicer arietinum</i>	XP_004510892.1
Euk. 8	<i>Vigna radiata</i> var. <i>radiata</i>	XP_014521222.1
Euk. 9	<i>Morus notabilis</i>	XP_010108055.1
Euk. 10	<i>Solanum tuberosum</i>	XP_015169639.1
Fun. 1	<i>Encephalitozoon cuniculi</i> GB-M1	NP_586424.1
Fun. 2	<i>Torulaspora delbrueckii</i>	XP_003680922.1
Fun. 3	<i>Trametes versicolor</i> FP-101664 SS1	XP_008044753.1
Fun. 4	<i>Kazachstania africana</i> CBS 2517	XP_003954912.1
Fun. 5	<i>Dichomitus squalens</i> LYAD-421 SS1	XP_007365208.1
Fun. 6	<i>Puccinia graminis</i> f. sp. <i>tritici</i> CRL 75-36-700-3	XP_003322083.2
Fun. 7	<i>Coniophora puteana</i> RWD-64-598 SS2	XP_007775552.1
Fun. 8	<i>Eremothecium cymbalariae</i> DBVPG#7215	XP_003647557.1
Fun. 9	<i>Moniliophthora roreri</i> MCA 2997	XP_007853138.1
Fun. 10	<i>Encephalitozoon intestinalis</i> ATCC 50506	XP_003073966.1
Met. 1	<i>Jaculus jaculus</i>	XP_004666971.1
Met. 2	<i>Chrysochloris asiatica</i>	XP_006863382.1
Met. 3	<i>Odobenus rosmarus divergens</i>	XP_004398631.1
Met. 4	<i>Priapulus caudatus</i>	XP_014664176.1
Met. 5	<i>Ovis aries musimon</i>	XP_011978134.1
Met. 6	<i>Trichechus manatus latirostris</i>	XP_004376138.1
Met. 7	<i>Dasypus novemcinctus</i>	XP_004447311.1
Met. 8	<i>Nannospalax galili</i>	XP_008854136.1, XP_008854137.1, XP_008854138.1
Met. 9	<i>Rattus norvegicus</i>	XP_008766000.1
Met. 10	<i>Capra hircus</i>	XP_005693518.1

Table S3B. Taxa details from Figure 2(C), (D). Labels indicate the taxa labels in Figures 2(C) and 2(D). Each label specifies the species name and accession number downloaded from RefSeq.

Label	Species	Accession no.
Bac. 1	<i>Leptospira meyeri</i>	WP_004787080.1
Bac. 2	<i>Nonomuraea coxensis</i>	WP_026214713.1
Bac. 3	<i>Actinomadura rifamycini</i>	WP_051300306.1
Bac. 4	<i>Marmoricola aequoreus</i>	WP_030484673.1
Bac. 5	<i>Kytococcus sedentarius</i>	WP_049758670.1
Bac. 6	<i>Nocardioides</i>	WP_056707204.1
Bac. 7	<i>Myxococcus fulvus</i>	WP_046713442.1
Bac. 8	<i>Chondromyces apiculatus</i>	WP_044234766.1
Bac. 9	<i>Phycococcus</i> sp. Root101	WP_056918311.1
Bac. 10	<i>Nitriliruptor alkaliphilus</i>	WP_052668139.1
Euk. 1	<i>Acanthamoeba castellanii</i> str. Neff	XP_004367908.1
Euk. 2	<i>Capsaspora owczarzaki</i> ATCC 30864	XP_004343108.1
Euk. 3	<i>Aphanomyces invadans</i>	XP_008862617.1
Euk. 4	<i>Saprolegnia diclina</i> VS20	XP_008605694.1
Euk. 5	<i>Acytostelium subglobosum</i> LB1	XP_012754808.1
Euk. 6	<i>Monosiga brevicollis</i> MX1	XP_001745020.1
Euk. 7	<i>Dictyostelium discoideum</i> AX4	XP_645447.1
Euk. 8	<i>Aureococcus anophagefferens</i>	XP_009039197.1
Euk. 9	<i>Monoraphidium neglectum</i>	XP_013896596.1
Euk. 10	<i>Fonticula alba</i>	XP_009493401.1
Fun. 1	<i>Punctularia strigosozonata</i>	XP_007381255.1
Fun. 2	<i>Wallemia ichthyophaga</i> EXF-994	XP_009266318.1
Fun. 3	<i>Coniophora puteana</i> RWD-64-598 SS2	XP_007766588.1
Fun. 4	<i>Puccinia graminis</i> f. sp. tritici CRL 75-36-700-3	XP_003327856.1
Met. 2	<i>Amyelois transitella</i>	XP_013199717.1
Met. 3	<i>Plutella xylostella</i>	XP_011563596.1
Met. 5	<i>Papilio polytes</i>	XP_013143494.1
Met. 6	<i>Bombyx mori</i>	XP_012546269.1

Table S3C. Taxa details from Figure 2(E), (F). Labels indicate the taxa labels in Figures 2(E) and 2(F). Each label specifies the species name and accession number downloaded from RefSeq.

*Note: All downloaded RefSeq sequences used for the analyses have been uploaded to the accompanying GitHub site.