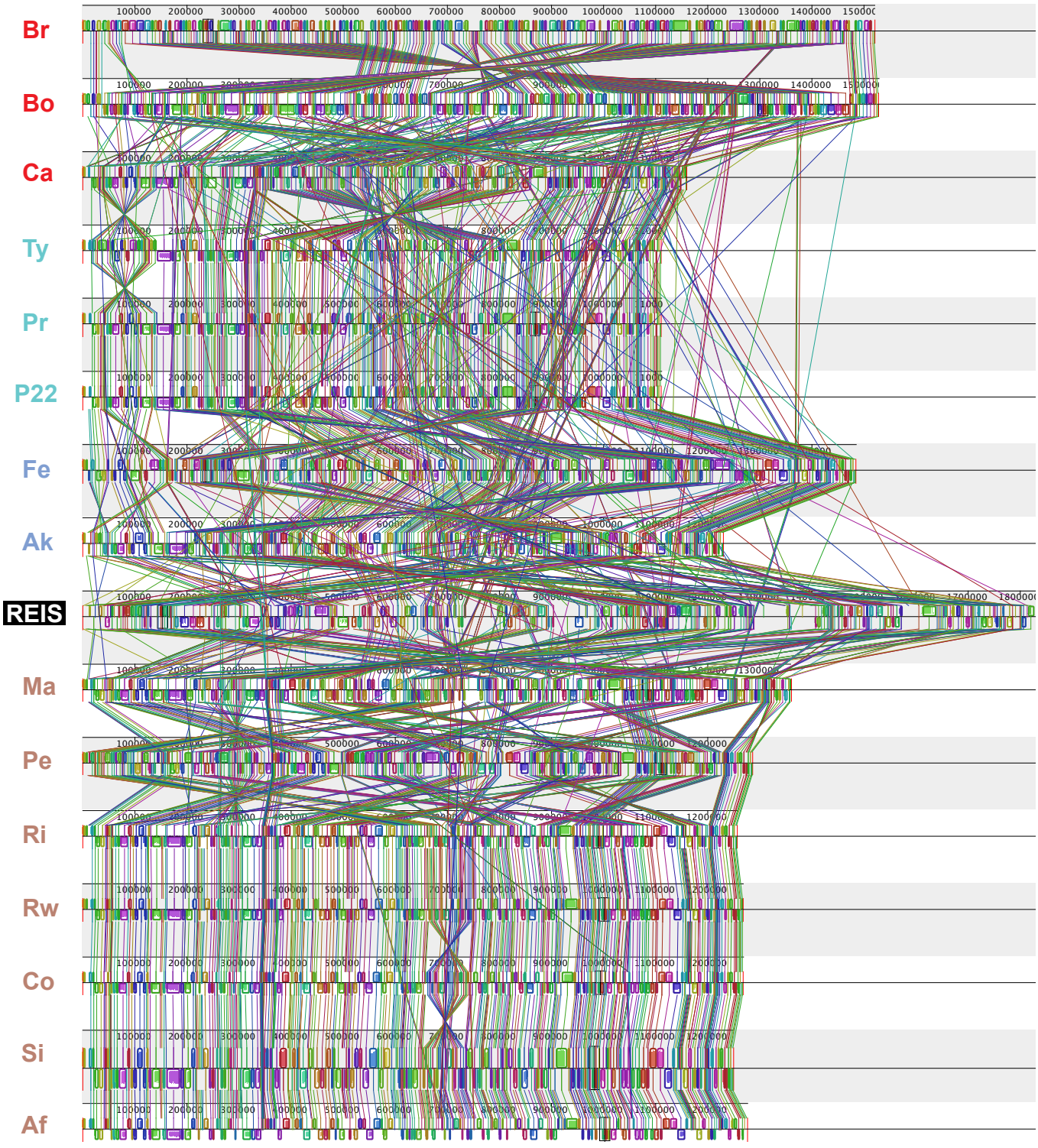
Fig. S1. Synteny analysis of 16 *Rickettsia* spp. genomes. The following genomes were included in all or some of the analyses: Br = *R. bellii* str. RML369-C (NC_007940), Bo = *R. bellii* str. OSU 85 389 (NZ_AARC01000001), Ca = *R. canadensis* str. McKiel (NZ_AAFF01000001), Ty = *R. typhi* str. Wilmington (NC_006142), Pr = *R. prowazekii* str. Madrid E (NC_000963), P22 = *R. prowazekii* str. P22 (CP001584), Fe = *R. felis* str. URRWXCal2 (NC_007109), Ak = *R. akari* str. Hartford (NZ_AAFE01000001), REIS = *Rickettsia* endosymbiont of *Ixodes scapularis*, Ma = *R. massilae* str. MTU5 (AAVR01000001), Pe = *R. peacockii* str. Rustic (NC_012730), Ri = *R. rickettsii* str. Sheila Smith (NZ_AADJ01000001), Rw = *R. rickettsii* str. Iowa (NC_010263), Co = *R. conorii* str. Malish 7 (NC_003103), Si = *R. sibirica* str. 246 (NZ_AABW01000001), Af = *R. africae* str. ESF-5 (AAUY01000001). Genome sequence alignments were performed using Mauve v.2.3.1 [1]. Unmodified Fasta files for each rickettsial genome were used as input, except that the *R. sibirica* genome sequence was reindexed using the reverse-complement of its circular permutation from the original position 668301, as previously analyzed [2]. (A) Alignment of 16 *Rickettsia* spp. genome sequences. (B) Alignment of 15 *Rickettsia* spp. genome sequences (excluding Pe). (C) Alignment of 15 *Rickettsia* spp. genome sequences (excluding Pe, switching positions of REIS and Ma). (D) Alignment of 14 *Rickettsia* spp. genome sequences (excluding Fe and Pe).

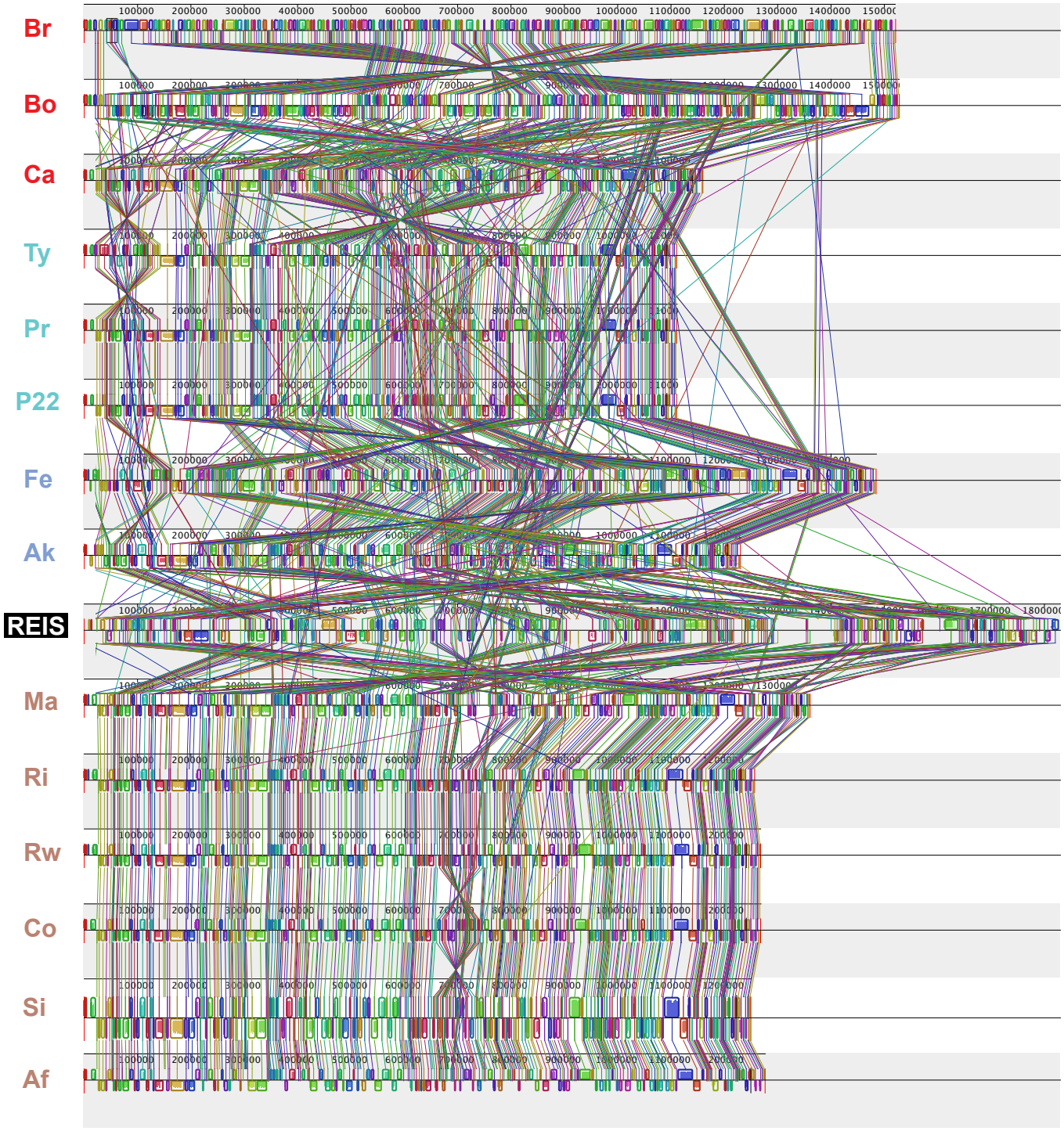
1. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.* PLoS One, 2010. **5**(6): p. e11147.
2. Gillespie, J.J., et al., *Rickettsia Phylogenomics: Unwinding the Intricacies of Obligate Intracellular Life.* PLoS ONE, 2008. **3**(4): p. e2018.

A



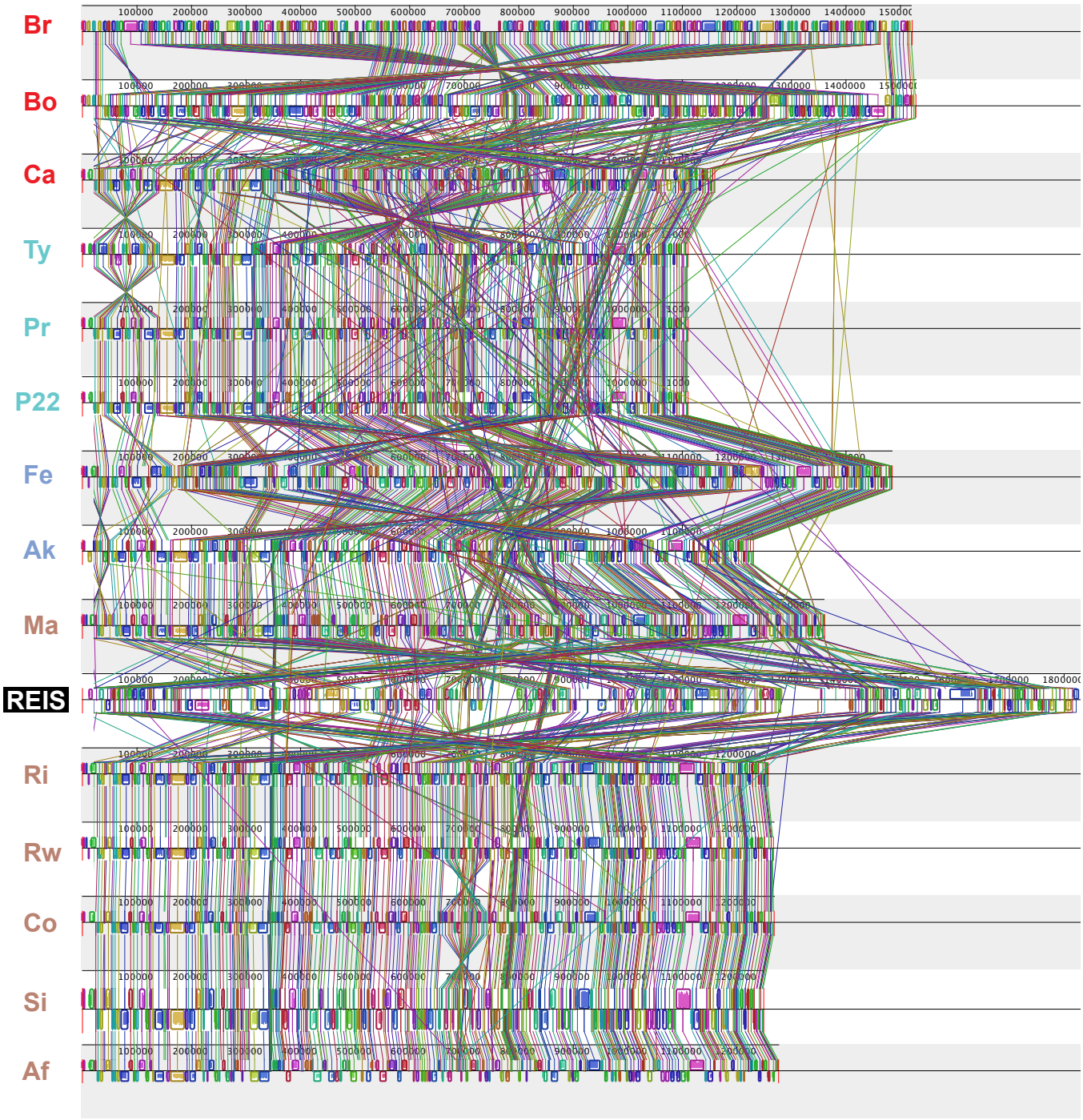
Alignment of 16 *Rickettsia* spp. genome sequences. The taxa are arranged according to the species phylogeny. The *R. bellii* genomes differ from one another by one large rearrangement, and little synteny exists across either *R. bellii* genome and *R. canadensis* (Ca). Ca is highly conserved in synteny with TG rickettsiae. Of the remaining derived genomes, *R. felis* (Fe), REIS and *R. peacockii* (Pe) have genome rearrangements that perturb an otherise highly conserved gene order. Further alingments (B-D) illustrate this with removal or repositioning of Fe, REIS and Pe.

B



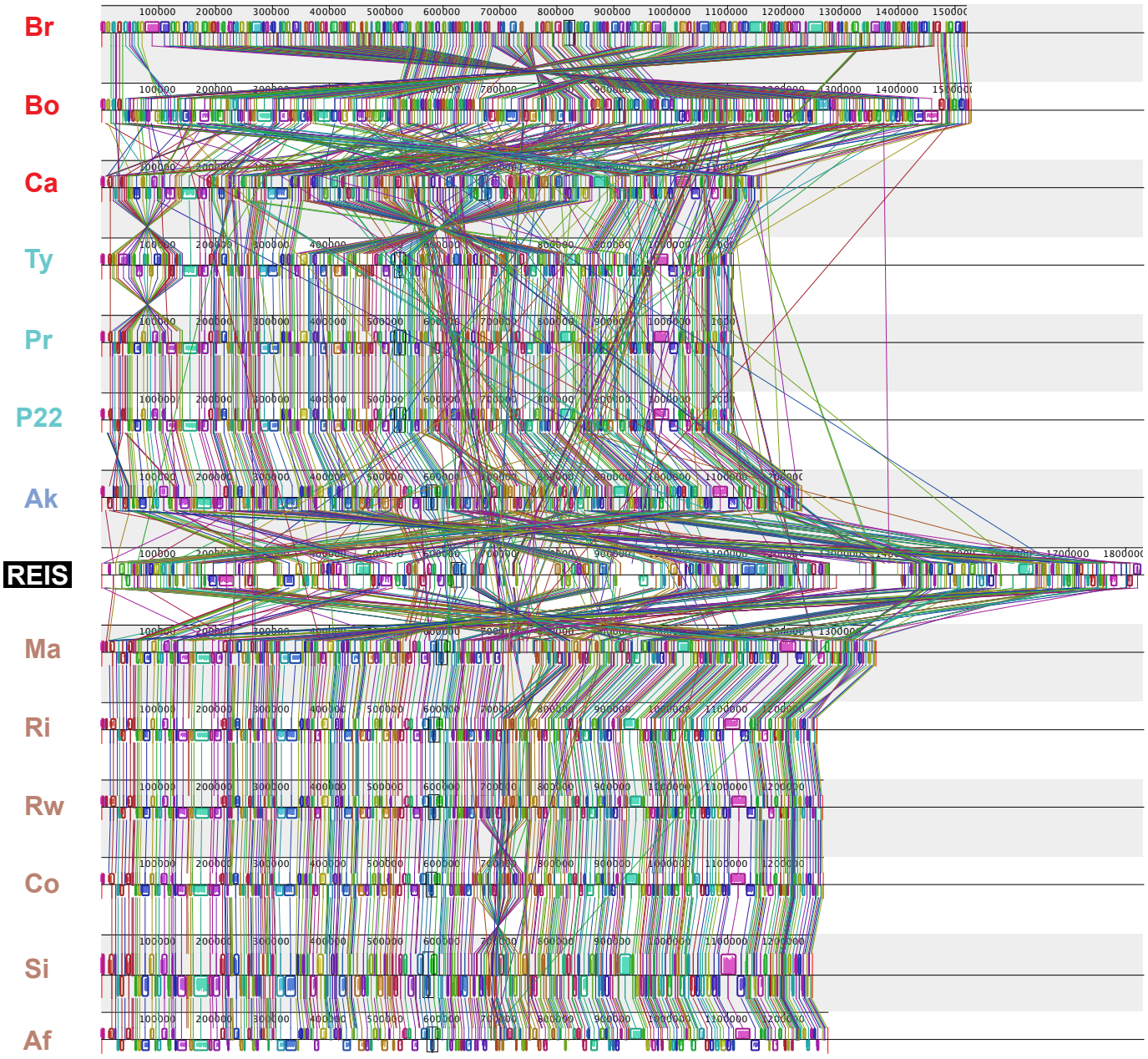
Alignment of 15 *Rickettsia* spp. genome sequences. The taxa are arranged according to the species phylogeny. *R. peacockii* (Pe) was not included in this alignment, as to better demonstrate the lack of synteny between REIS and the derived SFG rickettsiae.

C



Alignment of 15 *Rickettsia* spp. genome sequences. The taxa are arranged according to the species phylogeny. *R. peacockii* (Pe) was not included in this alignment, as to better demonstratethe lack of synteny between REIS and the derived SFG rickettsiae. REIS was switched in relation to *R. massiliae* (Ma) to further illustrate its larger size and position of rearrangements in relation to other SFG rickettsiae.

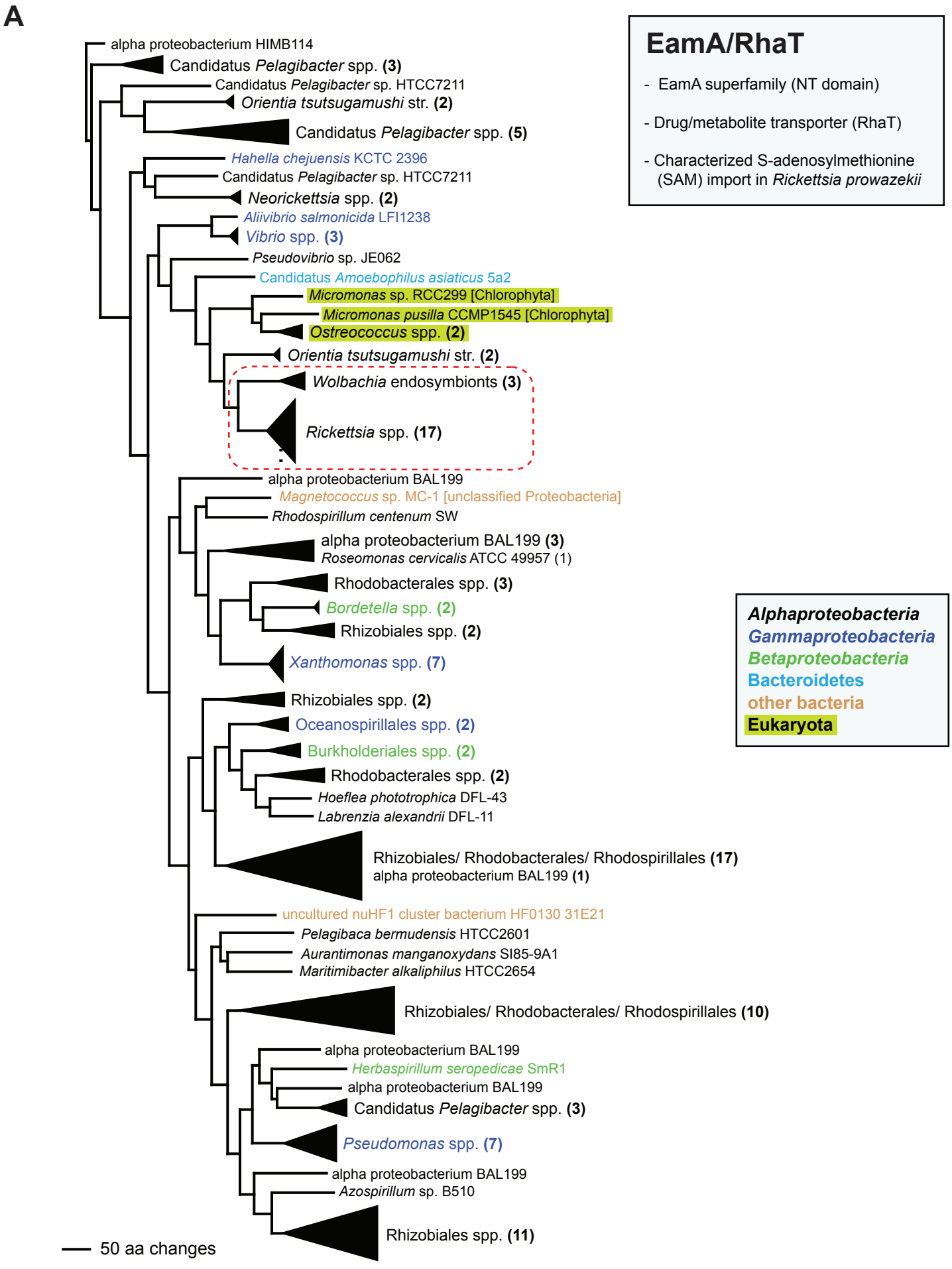
D

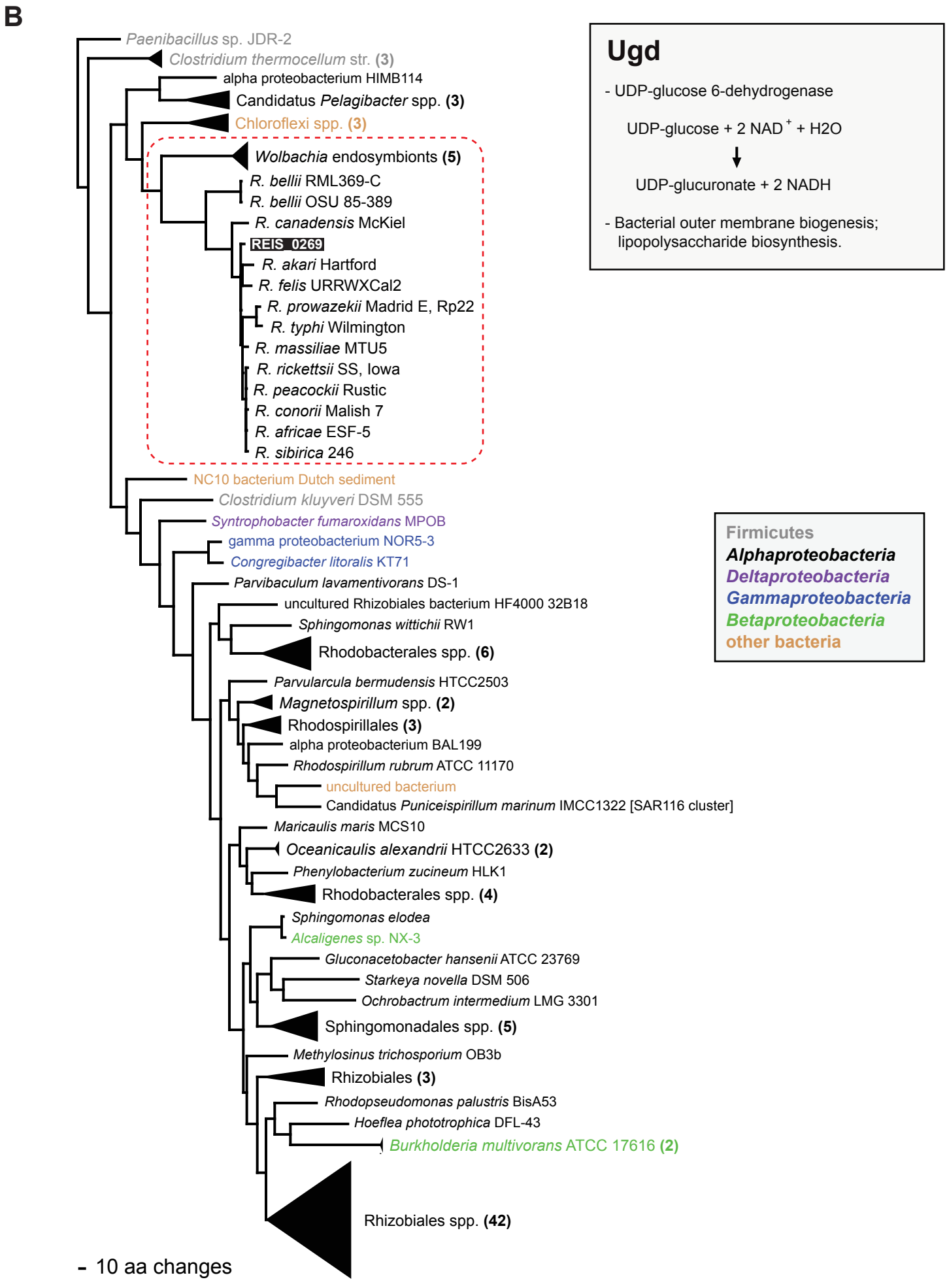


Alignment of 14 *Rickettsia* spp. genome sequences (excluding Fe and Pe). The taxa are arranged according to the species phylogeny. *R. felis* (Fe) and *R. peacockii* (Pe) were not included in this alignment, as to better demonstrate the lack of synteny between REIS and all of the *Rickettsia* spp. genomes except the *R. bellii* strains.

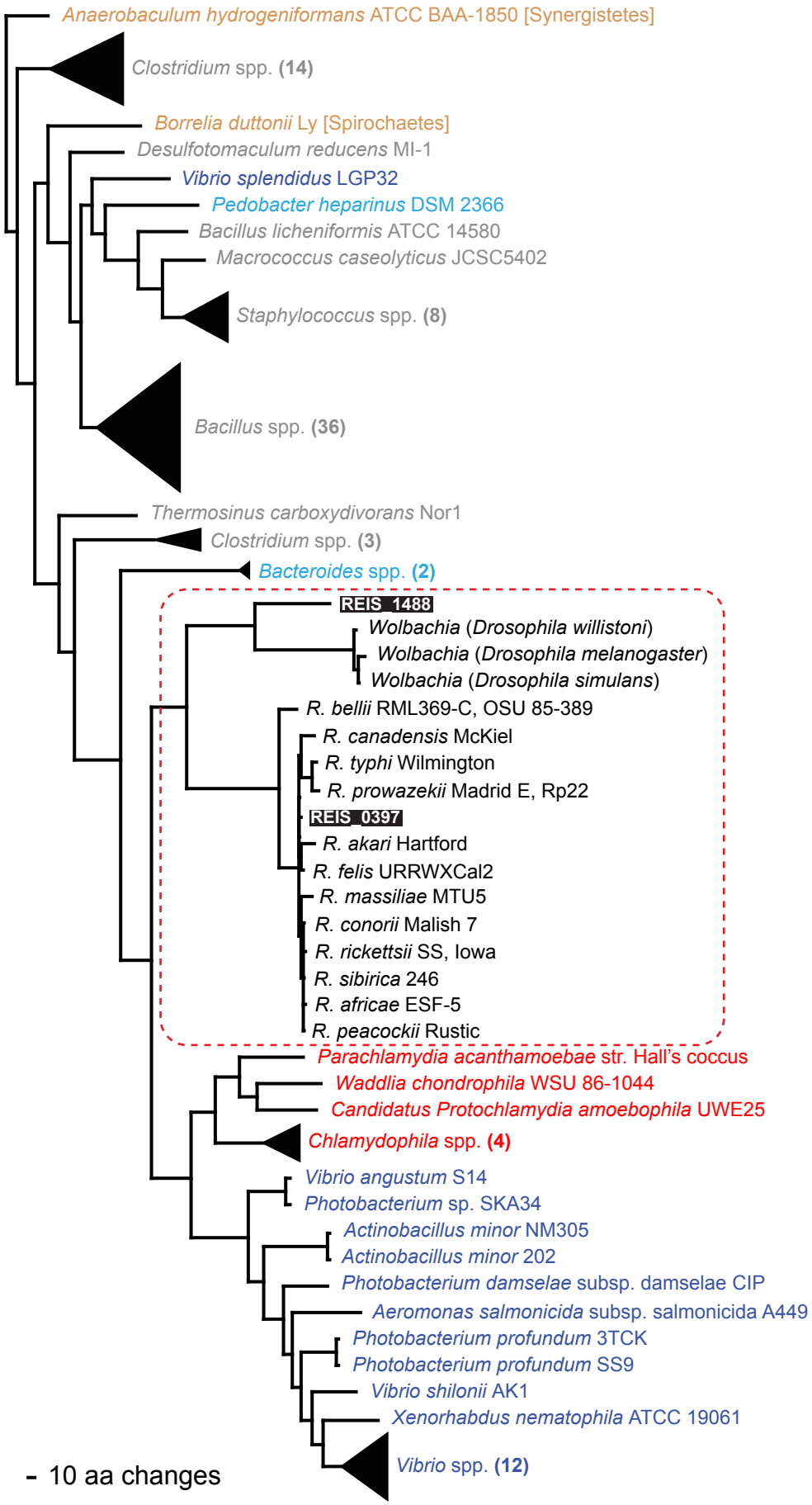
Fig. S2. Characteristics of the REIS genes with similarities to the WO-B prophage of *Wolbachia* spp. genomes. Nine of the 10 ORFs depicted in Fig. 2 are further illustrated here. See Fig. 6 for analysis of the ATP-binding multidrug resistance transporter MdlB. For each analysis, top blastp subjects (cut-off of 100) with significant alignments to the REIS queries were downloaded from NCBI and aligned using MUSCLE v3.6 [1, 2] (default parameters). For analyses other than the fusion proteins (KWG-OMeT and GT1-SAM) phylogenetic trees were estimated in PAUP* v4.0b10 (Altivec) under parsimony [3]. Majority rule consensus trees were constructed for analyses generating multiple equally parsimonious trees. (A) EamA, S-adenosylmethionine (SAM) transporter; (B) Ugd, UDP-glucose 6-dehydrogenase; (C) GlpT, glycerol-3-phosphate transporter; (D) LtaE, low specificity L-threonine aldolases; (E) PhyH, phytanoyl-CoA dioxygenase; (F) KWG-OMeT, N-terminal KWG-repeat domain fused to C-terminal O-methyltransferase (type 2) domain; (G) GT1-SAM, N-terminal glycosyltransferase (type1) domain fused to C-terminal radical SAM domain; (H) WcaG, nucleoside-diphosphate-sugar epimerase.

1. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
2. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
3. Swofford, D., *PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4 ed.* , 1999, Sinauer: Sunderland, MA.


A
EamA/RhaT
- EamA superfamily (NT domain)
- Drug/metabolite transporter (RhaT)
- Characterized S-adenosylmethionine (SAM) import in Rickettsia prowazekii
alpha proteobacterium HIMB114
Candidatus Pelagibacter spp. (3)
Candidatus Pelagibacter sp. HTCC7211
Orientia tsutsugamushi str. (2)
Candidatus Pelagibacter spp. (5)
Hahella chejuensis KCTC 2396
Candidatus Pelagibacter sp. HTCC7211
Neorickettsia spp. (2)
Aliivibrio salmonicida LFI1238
Vibrio spp. (3)
Pseudovibrio sp. JE062
Candidatus Amoebophilus asiaticus 5a2
Micromonas sp. RCC299 [Chlorophyta]
Micromonas pusilla CCMP1545 [Chlorophyta]
Ostreococcus spp. (2)
Orientia tsutsugamushi str. (2)
Wolbachia endosymbionts (3)
Rickettsia spp. (17)
alpha proteobacterium BAL199
Magnetococcus sp. MC-1 [unclassified Proteobacteria]
Rhodospirillum centenum SW
alpha proteobacterium BAL199 (3)
Roseomonas cervicalis ATCC 49957 (1)
Rhodobacterales spp. (3)
Bordetella spp. (2)
Rhizobiales spp. (2)
Xanthomonas spp. (7)
Rhizobiales spp. (2)
Oceanospirillales spp. (2)
Burkholderiales spp. (2)
Rhodobacterales spp. (2)
Hoeflea phototrophica DFL-43
Labrenzia alexandrii DFL-11
Rhizobiales/ Rhodobacterales/ Rhodospirillales (17)
alpha proteobacterium BAL199 (1)
uncultured nuHF1 cluster bacterium HF0130 31E21
Pelagibaca bermudensis HTCC2601
Aurantimonas manganoxydans SI85-9A1
Maritimibacter alkaliphilus HTCC2654
Rhizobiales/ Rhodobacterales/ Rhodospirillales (10)
alpha proteobacterium BAL199
Herbaspirillum seropedicae SmR1
alpha proteobacterium BAL199
Candidatus Pelagibacter spp. (3)
Pseudomonas spp. (7)
alpha proteobacterium BAL199
Azospirillum sp. B510
Rhizobiales spp. (11)
50 aa changes
Alphaproteobacteria
Gammaproteobacteria
Betaproteobacteria
Bacteroidetes
other bacteria
Eukaryota


B
Ugd
- UDP-glucose 6-dehydrogenase
UDP-glucose + 2 NAD+ + H2O
UDP-glucuronate + 2 NADH
- Bacterial outer membrane biogenesis; lipopolysaccharide biosynthesis.
Firmicutes
Alphaproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Betaproteobacteria
other bacteria
Paenibacillus sp. JDR-2
Clostridium thermocellum str. (3)
alpha proteobacterium HIMB114
Candidatus Pelagibacter spp. (3)
Chloroflexi spp. (3)
Wolbachia endosymbionts (5)
R. bellii RML369-C
R. bellii OSU 85-389
R. canadensis McKiel
REIS_0269
R. akari Hartford
R. felis URRWXCal2
R. prowazekii Madrid E, Rp22
R. typhi Wilmington
R. massiliae MTU5
R. rickettsii SS, Iowa
R. peacockii Rustic
R. conorii Malish 7
R. africae ESF-5
R. sibirica 246
NC10 bacterium Dutch sediment
Clostridium kluyveri DSM 555
Syntrophobacter fumaroxidans MPOB
gamma proteobacterium NOR5-3
Congregibacter litoralis KT71
Parvibaculum lavamentivorans DS-1
uncultured Rhizobiales bacterium HF4000 32B18
Sphingomonas wittichii RW1
Rhodobacterales spp. (6)
Parvularcula bermudensis HTCC2503
Magnetospirillum spp. (2)
Rhodospirillales (3)
alpha proteobacterium BAL199
Rhodospirillum rubrum ATCC 11170
uncultured bacterium
Candidatus Puniceispirillum marinum IMCC1322 [SAR116 cluster]
Maricaulis maris MCS10
Oceanicaulis alexandrii HTCC2633 (2)
Phenylobacterium zucineum HLK1
Rhodobacterales spp. (4)
Sphingomonas elodea
Alcaligenes sp. NX-3
Gluconacetobacter hansenii ATCC 23769
Starkeya novella DSM 506
Ochrobactrum intermedium LMG 3301
Sphingomonadales spp. (5)
Methylosinus trichosporium OB3b
Rhizobiales (3)
Rhodopseudomonas palustris BisA53
Hoeflea phototrophica DFL-43
Burkholderia multivorans ATCC 17616 (2)
Rhizobiales spp. (42)
- 10 aa changes

C

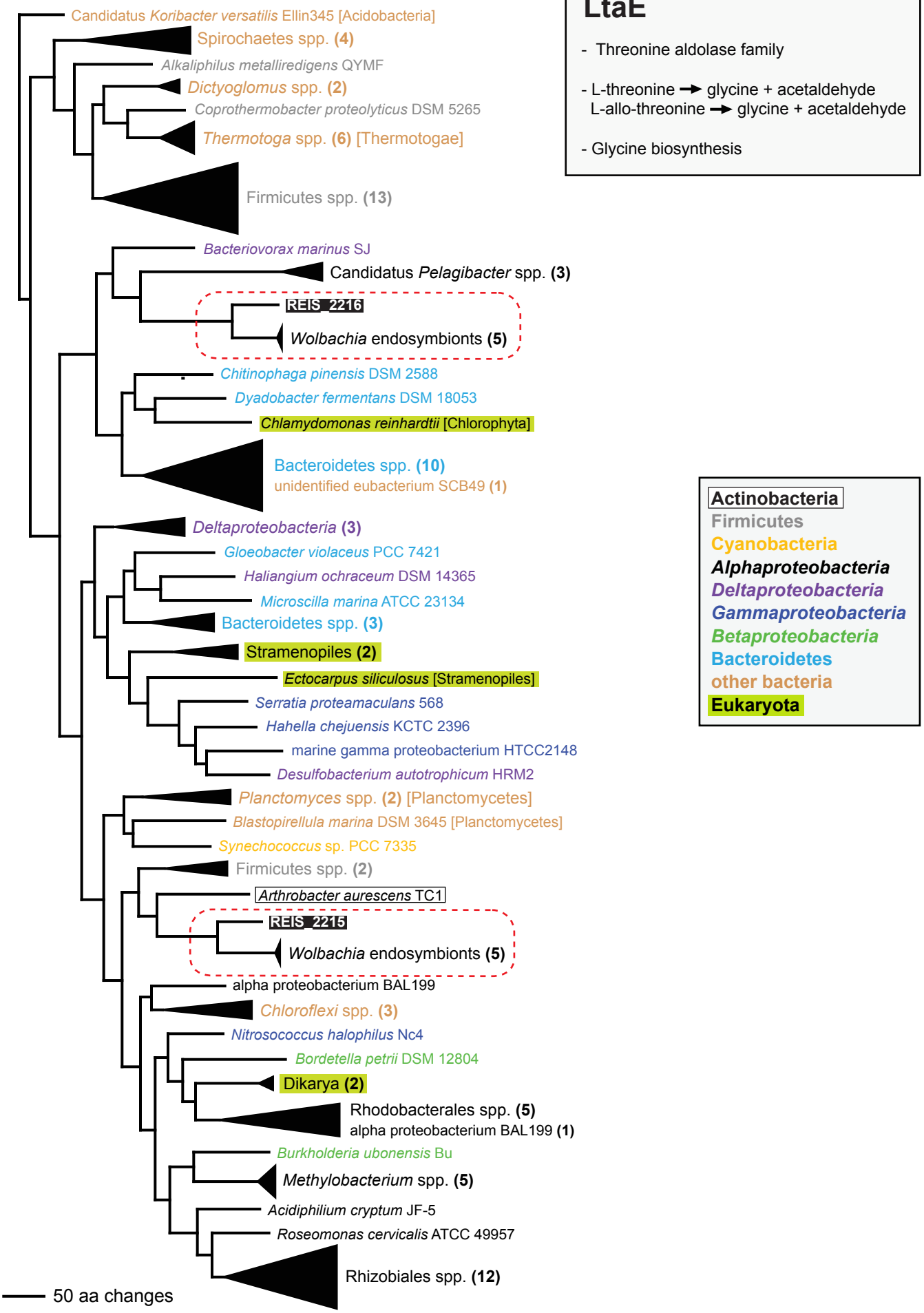


GlpT

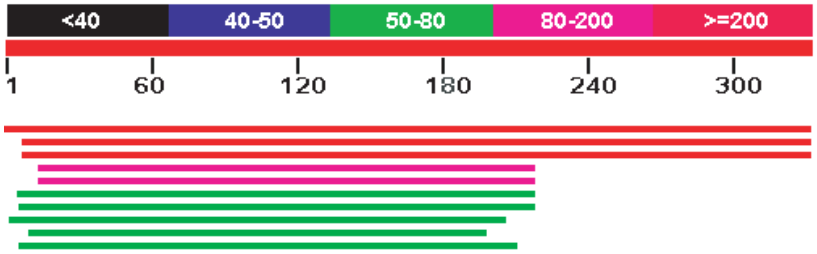
- Glycerol-3-phosphate transporter
- Host G3P exchanged for bacterial cytosolic phosphate
- G3P import has been reported in *Rickettsia prowazekii*, although a specific transporter has not been characterized.

Firmicutes
Chlamydiae
Alphaproteobacteria
Gammaproteobacteria
Bacteroidetes
other bacteria

D


LtaE
- Threonine aldolase family
- L-threonine → glycine + acetaldehyde
L-allo-threonine → glycine + acetaldehyde
- Glycine biosynthesis
Candidatus Koribacter versatilis Ellin345 [Acidobacteria]
Spirochaetes spp. (4)
Alkaliphilus metalliredigens QYMF
Dictyoglomus spp. (2)
Coprothermobacter proteolyticus DSM 5265
Thermotoga spp. (6) [Thermotogae]
Firmicutes spp. (13)
Bacteriovorax marinus SJ
Candidatus Pelagibacter spp. (3)
REIS_2216
Wolbachia endosymbionts (5)
Chitinophaga pinensis DSM 2588
Dyadobacter fermentans DSM 18053
Chlamydomonas reinhardtii [Chlorophyta]
Bacteroidetes spp. (10)
unidentified eubacterium SCB49 (1)
Deltaproteobacteria (3)
Gloeobacter violaceus PCC 7421
Haliangium ochraceum DSM 14365
Microscilla marina ATCC 23134
Bacteroidetes spp. (3)
Stramenopiles (2)
Ectocarpus siliculosus [Stramenopiles]
Serratia proteamaculans 568
Hahella chejuensis KCTC 2396
marine gamma proteobacterium HTCC2148
Desulfobacterium autotrophicum HRM2
Planctomyces spp. (2) [Planctomycetes]
Blastopirellula marina DSM 3645 [Planctomycetes]
Synechococcus sp. PCC 7335
Firmicutes spp. (2)
Arthrobacter aurescens TC1
REIS_2215
Wolbachia endosymbionts (5)
alpha proteobacterium BAL199
Chloroflexi spp. (3)
Nitrosococcus halophilus Nc4
Bordetella petrii DSM 12804
Dikarya (2)
Rhodobacterales spp. (5)
alpha proteobacterium BAL199 (1)
Burkholderia ubonensis Bu
Methylobacterium spp. (5)
Acidiphilium cryptum JF-5
Roseomonas cervicalis ATCC 49957
Rhizobiales spp. (12)
50 aa changes
Actinobacteria
Firmicutes
Cyanobacteria
Alphaproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Betaproteobacteria
Bacteroidetes
other bacteria
Eukaryota

E



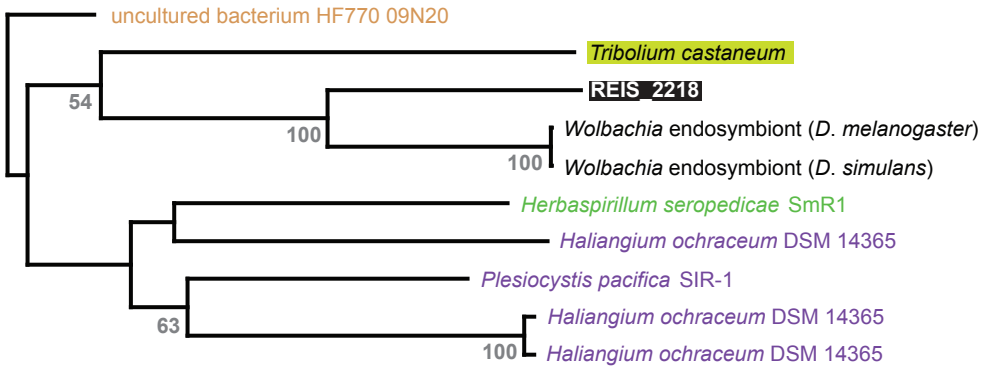
PhyH

- phytanoyl-CoA dioxygenase
- phytanoyl-CoA → 2-hydroxyphytanoyl-CoA
- lipid/fatty acid synthesis (eukaryotes)
- biosynthesis of mitomycin antibiotics/ polyketide fumonisin (bacteria)

```
uncultured bacterium HF770 09N20      6  VDQFRKDGFLVIGNLLDLE----TVGVLRQRFECLF-RGEFETGIS  (47)
Tribolium castaneum                  17  RQFYEDNGYIVIKNNVSHA----LLDEIQDRFIKICDGVADPGFMT  (59)
REIS                                  7  MEKLQDDGFFIIKNLISLD----LVTWHLNDIKNKIAGSAEEFGIS  (49)
Wolbachia endosymbiont (D. melanogaster)  ..  MKALQKNGFFIIKNLVPLE----LITQSLNDIISKISKLSQELGVS  (42)
Wolbachia endosymbiont (D. simulans)  ..  MKALQKNGFFIIKNLVPLE----LITQSLNDIISKISKLSQELGVS  (42)
Herbaspirillum seropedicae SmR1       9  VAFFREQGYLLLKGMVPADLRERMLAVTRDHLQRAVAPLEYEAEVG  (55)
Haliangium ochraceum DSM 14365        8  AQRFRDHGYFVLPGALSAD----DLALLRAACAWAVGVVDAAMDAA  (50)
Plesiocystis pacifica SIR-1           6  RRRYAELGWLVVPGLITRA----RALELARAFEVQTQRWAEAIDVP  (48)
Haliangium ochraceum DSM 14365       16  RSYFDVHGYAVIDTVLAVD----ERAALRGALEALWARCASEQSLP  (58)
Haliangium ochraceum DSM 14365       16  RSYFDVHGYAVIDTVLAVD----ERAALRGALEALWARCASEQSLP  (58)
                                             *
```

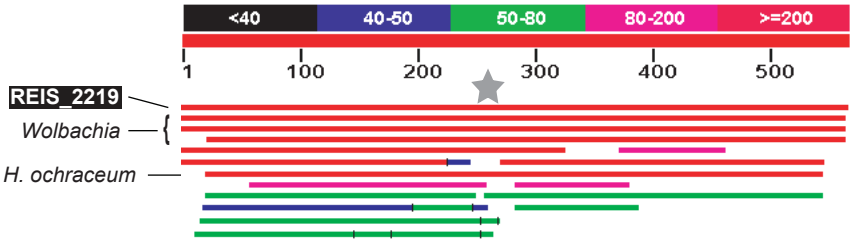
```
uncb   PDEV  18  WKADSTIARTVLR-QDLGREIAHLANWPGTRLF--QDNLLWKPP----GARPIGHHQDNAYVGWLMPQEIVSCWM  (137)
Tcas   VMKD  15  KIQDFLYDEVLFKYCSDKPVVDVIESIIGPNITGAHSMLINKPPDADPGASLHPLHQDLHYFPFRPADRIAASWT  (153)
REIS   VSDY   6  WVAPSQITRTISD-LLNDIIKDKLEKLLQCQIVNKKSNVICKTAD---LVDAVPFHQDISYNP-DNPYH-FSVWL  (128)
wMel   TSDY   6  WGTSSQVTRIVSQ-VLDKIIKNYLEKSLQCQILQNKSNVICKTAD---LIDAVPFHQDISY-SFNDPYH-FSVWL  (121)
wSim   TSDY   6  WGTSSQVTRIVSQ-VLDKIIKNYLEKSLQCQILQKKSNVICKTAD---LIDAVPFHQDISY-SFNDPYH-FSVWL  (121)
Hser   YEGA  18  WQRDPVYREWAGS-RALVAILHQLFDEPVCITLAHHNCVMTKHPA---FGTATGWHRDIRYWSFPRNEL-ISVWL  (147)
Hoch   GSER  15  HKRHPRLPGFLYS-PLMVALCRLALGRDAYLFL---EQFVVKGAG---DGAALPWHQDAGYLPFSSPPY-ITVWI  (136)
Ppac   PDEY  10  WERNASFKAQLFD-PRAAEIAAELIDCARVRVF--HDHLIAKPPL---GGGTIPWHRDLPNWPVAEPRA-LSCWL  (130)
Hoch   VTEY  10  WRHETTFAGFLAD-PRSWQTAALFMGQGGARLL--HDHVIAKPAG---GSGAVPWHQDQPYWPVDIDYG-VSCWC  (140)
Hoch   VTEY  10  WRHETTFAGFLAD-PRNWQTAALFMGQGGARLL--HDHVIAKPAG---GSGTVPWHQDQPYWPVDIDYG-VSCWC  (140)
                                                      *           * *             *
```

```
uncb   ALDDTQAQSGTLELIRGSHRWTHSQPDS  23  DIVSIEIPCGGASFHHGWVWHGSGNNRTASPRRALVLHGISSASEF   55  (289)
Tcas   AMERVDENNGCLYVIPGSHKW-ELYPHT  19  PKVNVVMEKGDTVFFHPILLHGSGPNRTKGFRKAISCHYADSNCYF   53  (298)
REIS   ALNDVYEASGALQIIKNSHNW-KIKPAV  18  QTITLPISAGEAIVFNSKLWHGSGENLNAKDRFAYVTRWVIKDKEF  114  (333)
wMel   ALNNVNKTSGALQVIEDSHNW-KIQPLV  18  KIKSLPISAGDAIVFDSRLWHGSDKNIDAKDRFAYVTRWVVKDKSL  117  (229)
wSim   ALNNVNKTSGALQVIEDSHNW-KIQPLV  18  KIKSLPISAGDAIVFDSRLWHGSDKNIDAKDRFAYVTRWVVKDKSL  117  (229)
Hser   ALGAETPENGALKFIPGSHKL-KLQPEQ  19  QGIALSLEPGDVVLFHSGLFHAAGRNDSDQVKCSAVFAYHG-----   21  (255)
Hoch   PLDDVDQDNGTLALLPYGR-A-GTRERV  20  ELVCAP--AGSVVLMSSTLLHRSGPNRSARPRRAFLAQY-------   46  (265)
Ppac   ALDDAPPDAGAMRFMPGGHRL-PETSSI  16  EAVPVPVSAGDAVFHHCLSWHCSPPNATRAWRRAYITIYLDADCTF  36  (255)
Hoch   PLEDVGPDGGCLEVIDGSHRW-GQSPPV  14  DRVELPLSAGGLVVLHSLTWHRSRSNRASGVRPAYITLWLPPDARY  167  (394)
Hoch   PLEDVGPDGGCLEVIDGSHRW-GQSPPV  14  DRVALPLSAGGLVVLHSLTWHRSRSNRASGVRPAYITLWLPPDARY  167  (394)
                 *                          *          *    *
```



Alphaproteobacteria
Deltaproteobacteria
Betaproteobacteria
other bacteria
Eukaryota

F

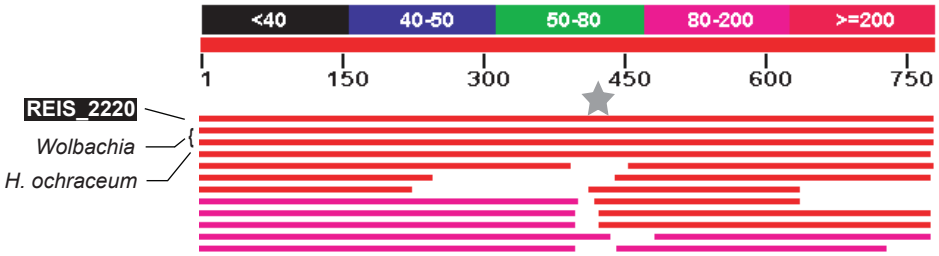


KWG-OMeT

- KWG-repeat domain
- O-methyltransferase domain
- protein fusion detected only in REIS, *Haliangium ochraceum*, and *Wolbachia* endosymbionts

```
Hoch  ------------------MSWREARVAATGTHHVCNGAPLYDERFDEVLKFHEPGLAPVRRGGRAWHIRSDGTAAYEQRFQRT  (65)
REIS  MQDMQYEKTVDQLTESNNLNSLKLIKISPCKQFHTLEEKPLYTTRFLHVEKFHEPGLAPVYDKTGAYHINLKGEAAYTKRFSKT  (84)
wMel  MQNMQHEKTISQLMKDS--EYLKSIRVSPCERFHQLSENPLYKNRFIRVDKFHEPGLAPVYDETGAYHIDAMGEAIYRDRFLKT  (82)
wSim  MQNMQHEKTISQLMKDS--EYLKSIRVSPCERFHQLSENPLYKNRFIRVDKFHEPGLAPVYDETGAYHIDAMGEAIYRDRFLKT  (82)
wRi   -------------MKDS--EYLKSIRVSPCERFHQLSENPLYKNRFIRVDKFHEPGLAPVYDETGAYHIDAMGEAIYRDRFLKT  (69)

Hoch  FGFYEGLAAVDSGVGWHHIHPDGRPLTATRYAWCGNFQNGYCAVRDHSGAYVHLTHAGAPAYGARWRYAGDFRDGVAVVQGADG  (149)
REIS  HGFYCGRAAVEDEARCYHIDSHANRVYKQSYQWVGNYQENICVVRKSN-KFYHIDLHGNKLYQEAYDYVGDFKDDIAVVY-KDG  (166)
wMel  FGFYCNRAAVEDDTGYYHIDPSGCRVYKHSYQWIGNYQEDICVIRKHD-KFFHIDLNGNRVYEQEYNYVGDFKDGIAVVH-KDG  (164)
wSim  FGFYCNRAAVEDDTGYYHIDPSGCRVYKHSYQWIGNYQEDICVIRKHD-KFFHIDLNGNRVYEQEYNYVGDFKDGIAVVH-KDG  (164)
wRi   FGFYCNRAAVEDDTGYYHIDPSGCRVYKHSYQWIGNYQEDICVIRKHD-KFFHIDLNGNRVYEQEYNYVGDFKDGIAVVH-KDG  (151)

Hoch  RSTHIDARGDFVHSVGFLDLDVFHKGFARARDEGGWMHVDMAGRAQYRRRFAAVEPFYNGQARVERFDGGLEVIDEAGDCVSTL  (233)
REIS  KSTHINPQGKLIHNKWYKQLGIFHKGFAIAEDNNGWFHVDIVGNAIYLQRYKTIEPFYNGLAMVKAYDDTLGQIDTTGNIKFVI  (250)
wMel  KATHINNHGKLVHNKWYKKLNVFHKGFAIAEDKHGWFHIDINGNPVYRQRFKMVEAFYNGMAKVETFEGVLGQIDITGNVKFSI  (248)
wSim  KATHINNHGKLVHNKWYKKLNVFHKGFAIAEDKHGWFHIDINGNPVYRQRFKMVEAFYNGMAKVETFEGVLGQIDITGNVKFSI  (248)
wRi   KATHINNHGKLVHNKWYKKLNVFHKGFAIAEDKHGWFHIDINGNPVYRQRFKMVEAFYNGMAKVETFEGVLGQIDITGNVKFSI  (235)

Hoch  RSP-LRSEFARLSEDMVGFWKTQTIHAAVALGVFEALPATDARIAEVCRLRPARARRLLRALAELGLL--ERGDDGWQASERGV  (314)
REIS  SLPELEAQIHKISAELAGFWKTYLINVAIELDLLNILPATTELLSKQLGIIEPNLQRLLRALWEIELISYNQNKDLWQVLPKGE  (334)
wMel  FDLDKESQVHKISAELSAFWKTYLANVAIELDLLNILPATMPVLSQKLNIIVPNLERLLRALWEIGFIDYDKNKDLWQLSSKGK  (332)
wSim  FDLDKESQVHKISAELSAFWKTYLANVAIELDLLNILPATMPVLSQKLNIIVPNLERLLRALWEIGFIDYDKNKDLWQLSSKGK  (332)
wRi   FDLDKESQVHKISAELSAFWKTYLANVAIELDLLNILPATMPVLSQKLNIIVPNLERLLRALWEIGFIDYDKNKDLWQLSSKGK  (319)

Hoch  YL-------QAAHPWTLAGAAREYGRDFSRMWEALPAALRADSDWCAPDIFGEVATDPERTAAHHQMLASYALHDYAGLPAALA  (391)
REIS  FLKNSSFLPQATKMWARVATEKN--------WLKITGLLKQKT-IYSFASFKEQETA--LKIEFYQALIGYTKLDTREFYNKIN  (407)
wMel  CFKEIPFLPKAAAMWARVAAEKN--------WLKIADILRQES-ISSFESFKEKETSEDRKIAFYQALLGYSRFDTKEFNSRVN  (407)
wSim  CFKEIPFLPKAAAMWARVAAEKN--------WLKIADILRQES-ISSFESFKEKETSEDRKIAFYQALLGYSRFDTKEFNSRVN  (407)
wRi   CFKEIPFLPKAAAMWARVAAEKN--------WLKIADILRQES-ISSFESFKEKETSEDRKIAFYQALLGYSRFDTKEFNSRVN  (394)

Hoch  LRGDERVIDAGGGLGTLAQLLV-AAYPRLRVVVLERAEVVALAEQAQDERVELCTADLFEPWGETGDAVVLARVLHDWDDAEAL  (474)
REIS  IADNTNILLF--GIHSLALIETLSNKDKNSINLNYYNDQEIPEELIKNYNVSLITQD--NL-AKNYDLSIFCRFLQNHDDEKVI  (486)
wMel  IDGAKNILLF--GVHSLFLAYS-DIHNKGSIGLYYYNEHKVPRQAVEDLKIKLITQE--ELSATNYELGVFCRFLQHYDDDKVL  (486)
wSim  IDGAKNILLF--GVHSLFLAYS-DIHNKGSIGLYYYNEHKVPRQAVEDLKIKLITQE--ELSATNYELGVFCRFLQHYDDNKVL  (486)
wRi   IDGAKNILLF--GVHSLFLAYS-DIHNKGSIGLYYYNEHKVPRQAVEDLKIKLITQE--ELSATNYELGVFCRFLQHYDDNKVL  (473)

Hoch  GLLRRARAALPPGGRVFIVEMLLSDSDMGGSLCDLHLLVATGGKERAEAEYAALLDEA-GFDCHGAQRFSSLSSILTGVAR  (554)
REIS  FYLRLAKDKKIT--RILLIETILTDNSPIGGAVDINIMVETGGKLRKLNDWEAILTQVGDLKISDVLPLTDYLSVIDIRYQ  (565)
wMel  SYLKLVKNKGIP--RVLVIETILDYYSPVGGSVDINVMVETGGKLRTLSDWKRILKQVKGLKVFSIVPLTDYLSVIDIRC-  (564)
wSim  SYLKLVKNKGIP--RVLVIETILDYYSPVGGSVDINVMVETGGKLRTLSDWKRILKQVKGLKIFSIVPLTDYLSVIDIRC-  (564)
wRi   SYLKLVKNKGIP--RVLVIETILDYYSPVGGSVDINVMVETGGKLRTLSDWKRILKQVKGLKIFSIVPLTDYLSVIDIRC-  (551)
```

G



GT1-SAM

- Glycosyltransferase domain
- Radical SAM domain
- protein fusion detected only in REIS, *Haliangium ochraceum*, and *Wolbachia* endosymbionts

```
Hoch   MKVLQVIHGYPMRYNAGSEVYTQTLCHGLADRHEVHVFTREEDSFAPDYRMRREHDPDDARITLHLVNNPRNKDRYR--AAGIDQRFA (86)
REIS   MKILKVIHGYPPYYRAGSEVYSQTLARKLSDNHEIQVFARHENSFLPSFHYSTVLDYGDPRILLHLINIPITKYRYKFINEEVNIRFE (88)
wMel   MKILKVIHGYPPYYSAGSEVYSQTLAHELANNNEVQIFTRHENSFLPDFHYTIALDCSDSRILLNLINIPTTKYRYKFINEEVNIKFK (88)
wSim   MKILKVIHGYPPYYSAGSEVYSQTLAHELANNNEVQIFTRHENSFLPDFHYTIALDCSDSRILLNLINIPTTKYRYKFINEEVNIKFK (88)
       ** * ******  * ****** ***   *    *   * * * ** *        *  * ** * * * *  * **         *

Hoch   ELLDRLRPEVVHVGHLNHLSTSLLREAATRSIPILYTLHDYWVMCPRGQFMQMFPEDGTDLWAACDGQDDRKCAERCYTRYFSGAPEE (174)
REIS   KIIDDFRPDLIHFGHLNHLSLNLPEIATKRGIPIVYTLHDFWLMCPRGRFIQ---RNSKELLQLCDGQEDKKCATECYKGYFTGDEGL (173)
wMel   RIIDNFQPDLIHFGHLNHLSITLPEVAFKENIPTIFTLHDFWLMCPRGRFIQ---RNSEDLLRLCDGQKDQKCATQCYKGHFTGDEEF (173)
wSim   RIIDNFQPDLIHFGHLNHLSITLPEVAFKENIPTIFTLHDFWLMCPRGRFIQ---RNSEDLLRLCDGQKDQKCATQCYKGHFTGDEEF (173)
          *   *   * *******  *   *    **   **** * ***** * *        *   **** * ***  ***  * *

Hoch   REHDLRYWTDWVARRMAHIREMADLVDVFIAPARYLRDRYRDAFGLPERKLVYLDYGFDRARLGGRKRTANEAFTFGYIGTHIPAKGI (262)
REIS   LNAEITYWQQWVSTRMKQTKKIVEYVDHFVAPSKFLMNKFVKEFNIPHTKISYLDYGFDLNRLKDKNRIPEEDFIFGYIGTHTPEKGI (261)
wMel   LNSDLNYWEQWVATRMKQTKKVVDYIDYFIAPSKFLMDKFTQDFNVPINKISYLDYGFDLNRLKNRNRVQEKGFIFGYIGTHTPEKGV (261)
wSim   LNSDLNYWEQWVATRMKQTKKVVDYIDYFIAPSKFLMDKFTQDFNVPINKISYLDYGFDLNRLKNRNRVQEKGFIFGYIGTHTPEKGV (261)
             **  **  **          * * **   *       *  *  ** *******  **    *     * ******* * **

Hoch   QVLLRAFGELRGSPRLRIWGRPRGQETAALKALAADLPGDAADRVEWLSEYRNQDIVRDVFDRTDAIVVPSVWVENSPLVIHEAQQAR (350)
REIS   DLLLKAFSSLSAKAKLRIWGALN-QETAGLKAIANHFPQKVKDKIEWMGNYENENIVTEVFNKVDAIVVPSIWGENSPLVIHEAEQLR (348)
wMel   DLLLKAFSRLSSKAKLRIWGAAR-EETKTLKAIADQFPHAVKERIEWMGSYNNKNIVTDVFNKVDAIVVPSIWGENSPLVIHEAQQLS (348)
wSim   DLLLKAFSRLSSKAKLRIWGAAR-EETKTLKAIADQFPHAVKERIEWMGSYNNKNIVTDVFNKVDAIVVPSIWGENSPLVIHEAQQLS (348)
         ** **  *     *****     **  *****   *       **   * *  **  **   ******* * ********** *

Hoch   VPVIAADVGGMAEYVHHEINGLQFEHRCERSLAAQMQRFVDDPAWARGLGERGYAFSESGDIPDIDTQVDDIERLYEAALASRDSARV (438)
REIS   IPVITADYGGMAEYVQDGINGLLFKHRDIESLSEKMENLNIDKNLYTKLTKKGYLYSNDGNVPSISEHTMELEKIYFNIIANKKKS-V (435)
wMel   VPVITADYGGMAEYVRDGINGLLFKHRDTNSLSEKMQVLSTDQELYSELTQRGYLYTKNGNVPSISEHTEKLNKIYRNIIENKGKS-V (435)
wSim   VPVITADYGGMAEYVRDGINGLLFKHRDTNSLSEKMQVLSTDQELYSELTQRGYLYTKNGNVPSISEHTEKLNKIYRNVIENKGKS-V (435)
        *** ** *******   **** * **   **   *     *      *   **     *  * *         *         *

Hoch   EVGPGPWRITFDTNPDTCNMRCVMCEEHSPHSPLQTLRKAEGRARREMPIELIREVVADAAAHGLREIIPSTMGEPLLYEHFEAILAL (526)
REIS   STKPGPWRITFDTNPDYCNYACIMCECFSPYSKVKENKRAEGIKPKIMPIATIRKVIQEAAGTPLREIIPSTMGEPLMYKHFNEIINI (523)
wMel   ATKPGPWRITFDTNPDYCNFACVMCECFSPYSKVKEEKKAKGIKPKIMSIETIRKVIKEADGTPLREIIPSTMGEPLMYKSFDEIINL (523)
wSim   ATKPGPWRITFDTNPDYCNFACVMCECFSPYSKVKEEKKAKGIKPKIMSIETIRKVIKEADGTPLREIIPSTMGEPLMYKSFDEIINL (523)
          ************* **  * ***  ** *       * *     * *  ** * *   *    ************ *  *  *

Hoch   CVEHGVRLNLTTNGSFPRLGARAWAERIVPVTSDVKISWNGATKATQEAIMLGSDWEAVLDNVRAFLAVRDGHAARGGNRCRVTFQLT (614)
REIS   CHEYGLKLNLTTNGSFPAKGAQKWAELLVPILSDIKISWNGASKEVNEKIMIGSKWEAVTKNLKTFLDVRDEYFSKTKKRCSVTLQLT (611)
wMel   CHEFGLKLNLTTNGSFPIKGARKWAELLVPILSDVKISWNGATKETHERIMKGSKWEVVTENLKTFLEVRDKYFSDTGERCTVTLQLT (611)
wSim   CHEFGLKLNLTTNGSFPIKGARKWAELLVPILSDVKISWNGATKETHERIMKGSKWEVVTENLKTFLEVRDKYFSDTGERCTVTLQLT (611)
       * * *  **********  **  ***  **  ** ******* *   * ** ** ** * *  *   ** ***        ** ** ***

Hoch   FLEANVAELADIVRLAVSLGVDRVKGHHLWAHFDEIKEQSMRRSPEAIQRWNAAVLAAREAAAERPLPNGKYVLLENIFLLEEQATAD (702)
REIS   FLESNLLELYDIVKMAIGLGIDRVKGHHLWAHFEEIKDLSMRRDETSINRWNTEVKRLYELRDAMLLPNGKKIILENFTILAKEGVED (699)
wMel   FLESNLHELYDIVEMAIKNGIDRVKGHHLWAHFKEIKDLSMRRDELAISRWNTEVGRLYELRDNMLLPNGKKIKLENFTILSQEGIKD (699)
wSim   FLESNLHELYDIVEMAIKNGIDRVKGHHLWAHFKEIKDLSMRRDELAISRWNTEVGRLYELRDNMLLPNGKKIKLENFTILSQEGIKD (699)
       *** *  ** ***  *   * ************ ***  ****    * ***  *    *      *****   ***   *      *

Hoch   LAPGGPCPFLGKEAWVSAEGRFDPCCAPDAQRRTLGSFGNLGDSGIMEIWNGPAYRELAASYRNRALCLRCNMRKPAEEP       (782)
REIS   LAPGGPCPFLGKEAWVNPEGKFSPCCAPDELRKTLGNFGNVNEVKLEDIWQSNQYKDLQKNYFNHELCKTCNMRKPLIS-       (778)
wMel   LAPGGQCPFLGKEAWINNEGKFSPCCAPDELRKTLGDFGNVNEVKLEEIWQSSEYLNLQKNYLNYELCKTCNMRKPLVN-       (778)
wSim   LAPGGQCPFLGKEAWINNEGKFSPCCAPDELRKTLGDFGNVNEVKLEEIWQSSEYLNLQKNYLNYELCKTCNMRKPLVN-       (778)
       ***** *********   ** * ******  * *** ***         **    *  *   * *  **  ******
```

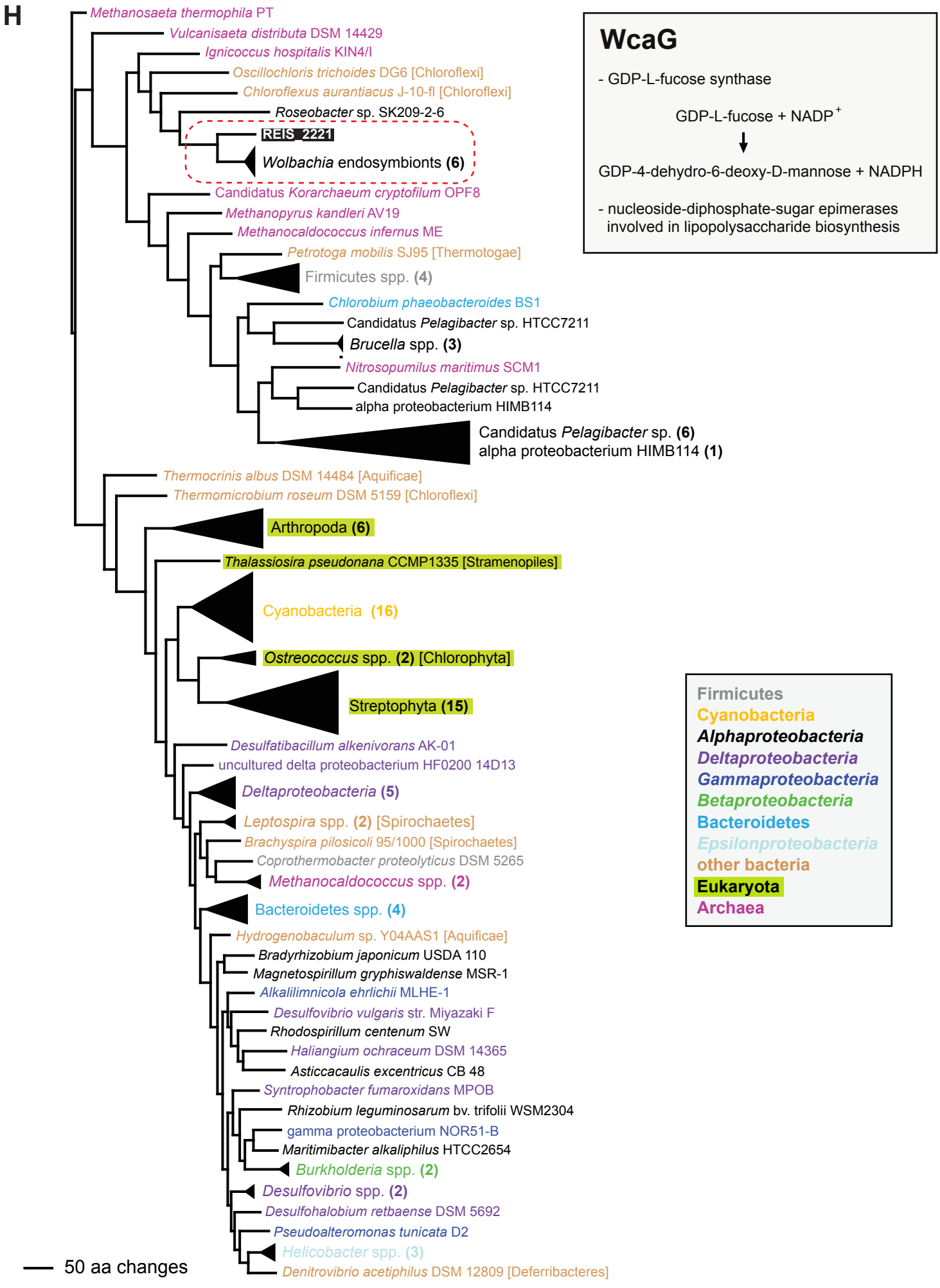

H
Methanosaeta thermophila PT
Vulcanisaeta distributa DSM 14429
Ignicoccus hospitalis KIN4/I
Oscillochloris trichoides DG6 [Chloroflexi]
Chloroflexus aurantiacus J-10-fl [Chloroflexi]
Roseobacter sp. SK209-2-6
REIS_2221
Wolbachia endosymbionts (6)
Candidatus Korarchaeum cryptofilum OPF8
Methanopyrus kandleri AV19
Methanocaldococcus infernus ME
Petrotoga mobilis SJ95 [Thermotogae]
Firmicutes spp. (4)
Chlorobium phaeobacteroides BS1
Candidatus Pelagibacter sp. HTCC7211
Brucella spp. (3)
Nitrosopumilus maritimus SCM1
Candidatus Pelagibacter sp. HTCC7211
alpha proteobacterium HIMB114
Candidatus Pelagibacter sp. (6)
alpha proteobacterium HIMB114 (1)
Thermocrinis albus DSM 14484 [Aquificae]
Thermomicrobium roseum DSM 5159 [Chloroflexi]
Arthropoda (6)
Thalassiosira pseudonana CCMP1335 [Stramenopiles]
Cyanobacteria (16)
Ostreococcus spp. (2) [Chlorophyta]
Streptophyta (15)
Desulfatibacillum alkenivorans AK-01
uncultured delta proteobacterium HF0200 14D13
Deltaproteobacteria (5)
Leptospira spp. (2) [Spirochaetes]
Brachyspira pilosicoli 95/1000 [Spirochaetes]
Coprothermobacter proteolyticus DSM 5265
Methanocaldococcus spp. (2)
Bacteroidetes spp. (4)
Hydrogenobaculum sp. Y04AAS1 [Aquificae]
Bradyrhizobium japonicum USDA 110
Magnetospirillum gryphiswaldense MSR-1
Alkalilimnicola ehrlichii MLHE-1
Desulfovibrio vulgaris str. Miyazaki F
Rhodospirillum centenum SW
Haliangium ochraceum DSM 14365
Asticcacaulis excentricus CB 48
Syntrophobacter fumaroxidans MPOB
Rhizobium leguminosarum bv. trifolii WSM2304
gamma proteobacterium NOR51-B
Maritimibacter alkaliphilus HTCC2654
Burkholderia spp. (2)
Desulfovibrio spp. (2)
Desulfohalobium retbaense DSM 5692
Pseudoalteromonas tunicata D2
Helicobacter spp. (3)
Denitrovibrio acetiphilus DSM 12809 [Deferribacteres]
50 aa changes
WcaG
- GDP-L-fucose synthase
GDP-L-fucose + NADP+
GDP-4-dehydro-6-deoxy-D-mannose + NADPH
- nucleoside-diphosphate-sugar epimerases involved in lipopolysaccharide biosynthesis
Firmicutes
Cyanobacteria
Alphaproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Betaproteobacteria
Bacteroidetes
Epsilonproteobacteria
other bacteria
Eukaryota
Archaea

Fig. S3. Generation of orthologous protein families across 16 Rickettsiaceae genomes. OrthoMCL [1] was used to generate orthologous groups (OGs) from a total of 20,035 predicted proteins across sixteen complete Rickettsiaceae genomes (summarized in gray box at top). Throughout the schema, blue and red numbers depict protein and OG counts, respectively. Characteristics of different protein family categories are described moving counterclockwise from the box at top (following the gray arrows). **Unique proteins** are subdivided into true singletons and dupletons, which are unique proteins that are duplicated within a genome. Unique proteins are further described for each genome (green inset), with a gray box illustrating the 32% of the REIS genome comprised of unique proteins. **Core proteins** are subdivided into perfect families, which have one protein from every genome, and imperfect, which have at least one genome contributing multiple proteins. The total Rickettsiaceae core genome is comprised of 468 OGs, with an additional 166 OGs present in all *Rickettsia* genomes (thus the core OGs of *Rickettsia* genomes total 634). **Non-conserved proteins**, which are encoded within more than one genome (but not all genomes), comprise 38.7% of all proteins across the 16 genomes. These are subdivided into singular OGs, which do not contain gene duplications, and multiple OGs, which do contain gene duplications. More non-conserved OGs do not include REIS proteins (**REIS -**, 60.4%), however the non-conserved OGs including REIS proteins (**REIS +**) have a greater number of duplicate proteins, indicative of the proliferated MGEs in some genomes (e.g., STG rickettsiae, REIS). REIS encodes 836 non-conserved proteins, which is far greater than any other Rickettsiaceae genome (see bar graphs, REIS is distinguished by the red dashed box). The majority of duplicate proteins encoded by the REIS genome are MGEs, especially TNPs. Importantly, the **REIS +** singular OGs have an average of 8.5 proteins, whereas the **REIS -** singular OGs only have an average of four proteins. This indicates that the protein families lacking an REIS protein are decaying more rapidly from the core Rickettsiaceae genome. Genome codes as follows: Bg, *O. tsutsugamushi* str. Boryong; Ik, *O. tsutsugamushi* str. Ikeda;Br, *R. bellii* str. RML369-C; Bo, *R. bellii* str. OSU 85 389; Ca, *R. canadensis* str. McKiel; Pr, *R. prowazekii* str. Madrid E; Ty, *R. typhi* str. Wilmington; Fe, *R. felis* str. URRWXCal2; Ak, *R. akari* str. Hartford; REIS, *Rickettsia* endosymbiont of *Ixodes scapularis*; Ma, *R. massilae* str. MTU5; Ri, *R. rickettsii* str. Sheila Smith; Rw, *R. rickettsii* str. Iowa; Co, *R. conorii* str. Malish 7; Si, *R. sibirica* str. 246; Af, *R. africae* str. ESF-5.

1. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

From a total of 20,035 proteins across 16 genomes, OrthoMCL grouped 18,035 proteins into 2,069 OGs. A subset of these (237 OGs) contains two or more proteins from only one genome (dupleton, D), totaling 1,063 proteins. True singletons (S), present only once in a single genome, total 1,350 proteins.

Unique
2,413 proteins are unique (U) to a single genome; 32% of the REIS genome encodes unique proteins.

	REIS	16 genomes
U	739; 64	2,413; 237
D	351; 64	1,063; 237
S	388; 0	1,350; 0

Core
7,521 proteins are core Rickettsiaceae; an additonal 2,343 proteins define the core *Rickettsia*.

perfect (P): one protein per genome
imperfect (I): one genome w/ 1 ≥ protein

	Rickettsiaceae	*Rickettsia*
P	7,335; 459	2,226; 159
I	186; 9	117; 7

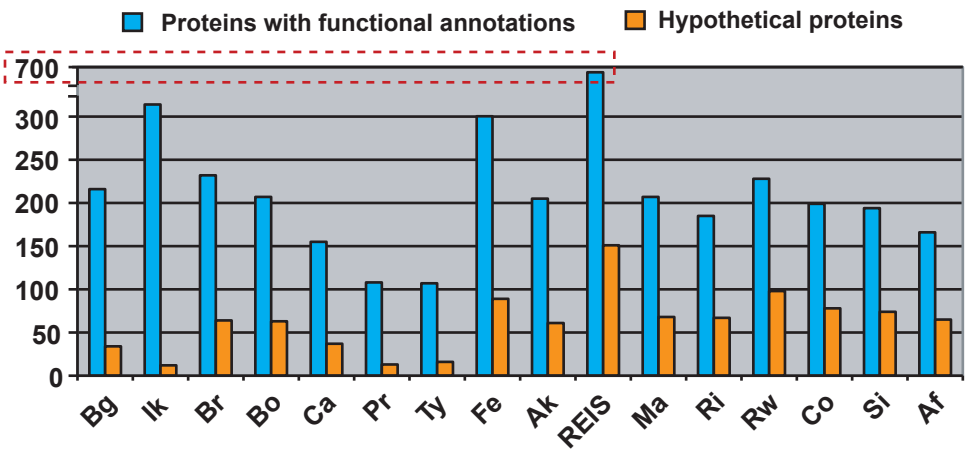
Non-conserved
7,758 proteins from 1198 OGs exist in 2-15 genomes. 60% of these OGs (723) lack REIS proteins.

singular (S): one protein per 2-15 genomes
multiple (M): ≥ 1 proteins per 2-15 genomes

	REIS +	REIS -
S	2,883; 338	2,470; 640
M	1,816; 137	589; 83

					% of genome
Bg	102	125; 31	227	19.2
Ik	328	529; 117	857	43.6
Br	27	2; 1	29	2.0
Bo	36	2; 1	38	2.6
Ca	40	0; 0	40	3.7
Pr	8	0; 0	8	1.0
Ty	7	2; 1	9	1.1
Fe	118	52; 22	170	11.2
Ak	46	0; 0	46	3.7
REIS	388	351; 64	739	32.0
Ma	50	0; 0	50	4.2
Ri	12	0; 0	12	0.9
Rw	144	0; 0	144	10.2
Co	13	0; 0	13	1.0
Si	9	0; 0	9	0.8
Af	22	0; 0	22	2.0

Non-conserved, REIS +
Of 4,699 proteins from 475 OGs, REIS encodes 836, or 18%. The avg. for the other genomes is 5%. A total of 423 of these REIS proteins are MGEs, dominated by TNPs (304). NOTE: the avg. no. of proteins in singular OGs (338) is 8.5.



Non-conserved, REIS -
REIS lacks homologs to 3,059 proteins (723 OGs). Proteins missing from the core Rickettsiales and core Rickettsia genomes are shown in Table S1. NOTE: the avg. no. of proteins in singular OGs (640) is 4.

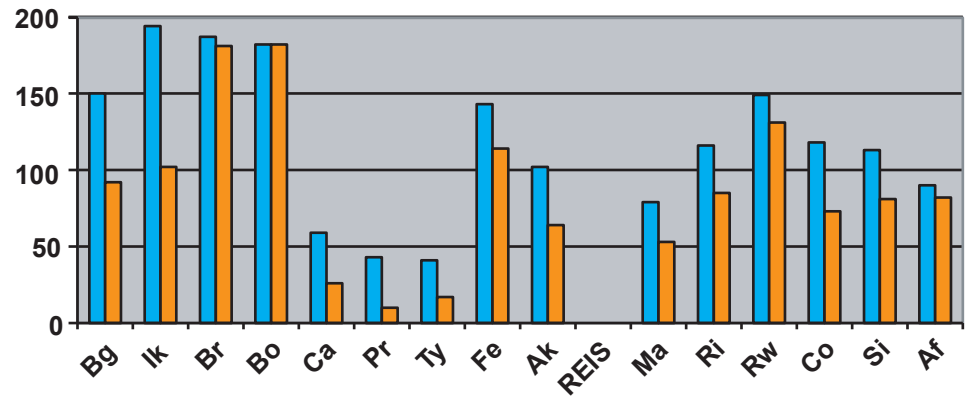


Fig. S4. Whole genome-based phylogeny estimation for 46 Rickettsiales taxa. An automated workflow for gene family selection and tree building was implemented through a set of Perl scripts [1]. All protein sequences annotated by RAST [2] for 46 Rickettsiales genomes (plus two outgroup taxa) were downloaded from PATRIC [3]. The following pipeline was implemented to estimate Rickettsiales phylogeny: BLAT (refined BLAST algorithm) [4] searches were performed to identify similar protein sequences between all genomes, including the two outgroup taxa. To predict initial homologous protein sets, mcl [5] was used to cluster BLAT results, with subsequent refinement of these sets using in-house hidden Markov models [6]. These protein families were then filtered to include only those with membership in >80% of the analyzed genomes (39 or more taxa included per protein family). Multiple sequence alignment of each protein family was performed using MUSCLE (default parameters) [7, 8], with masking of regions of poor alignment (length heterogeneous regions) done using Gblocks (default parameters) [9, 10]. All modified alignments were then concatenated into one dataset. Tree-building was performed using FastTree [11]. Support for generated lineages was estimated using a modified bootstrapping procedure, with 100 pseudoreplications sampling only half of the aligned protein sets per replication (NOTE: standard bootstrapping tends to produce inflated support values for very large alignments). All branches in the illustrated tree were supported by 100%. Local refinements to tree topology were attempted in instances where highly supported nodes have subnodes with low support. This refinement is executed by running the entire pipeline on only those genomes represented by the node being refined (with additional sister taxa for rooting purposes). The refined subtree was then spliced back into the full tree. More information pertaining to this phylogeny pipeline is available at PATRIC.

1. Williams, K.P., et al., *Phylogeny of gammaproteobacteria.* J Bacteriol, 2010. **192**(9): p. 2305-14.
2. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.
3. Snyder, E.E., et al., *PATRIC: the VBI PathoSystems Resource Integration Center.* Nucleic Acids Res, 2007. **35**(Database issue): p. D401-6.
4. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
5. Van Dongen, S., *Graph Clustering Via a Discrete Uncoupling Process.* SIAM. J. Matrix Anal. & Appl., 2008. **30**: p. 121-141.
6. Durbin, R., et al., *Biological sequence analysis: probabilistic models of proteins and nucleic acids*1998: Cambridge University Press.
7. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
8. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
9. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Mol Biol Evol, 2000. **17**(4): p. 540-52.
10. Talavera, G. and J. Castresana, *Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.* Syst Biol, 2007. **56**(4): p. 564-77.
11. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.

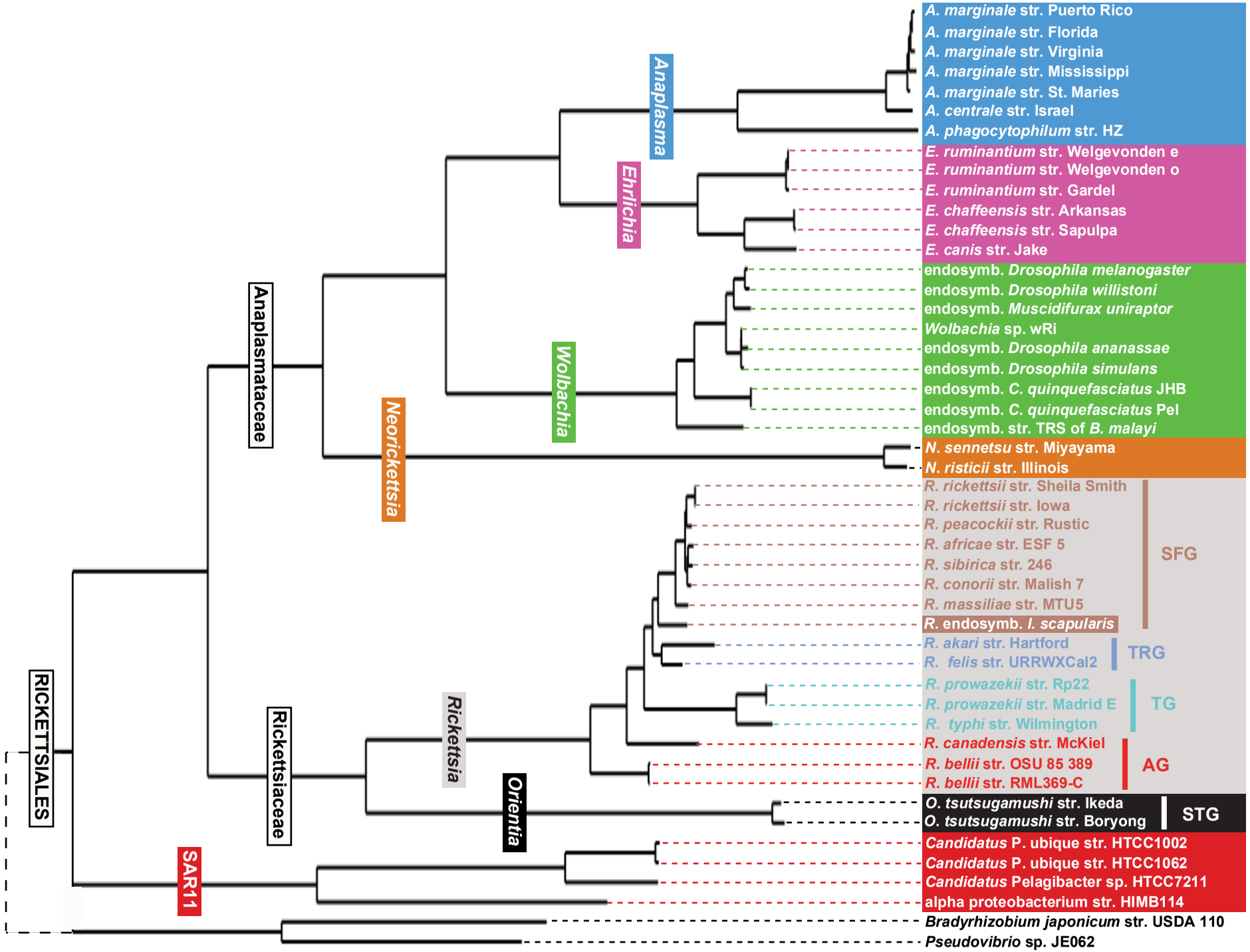
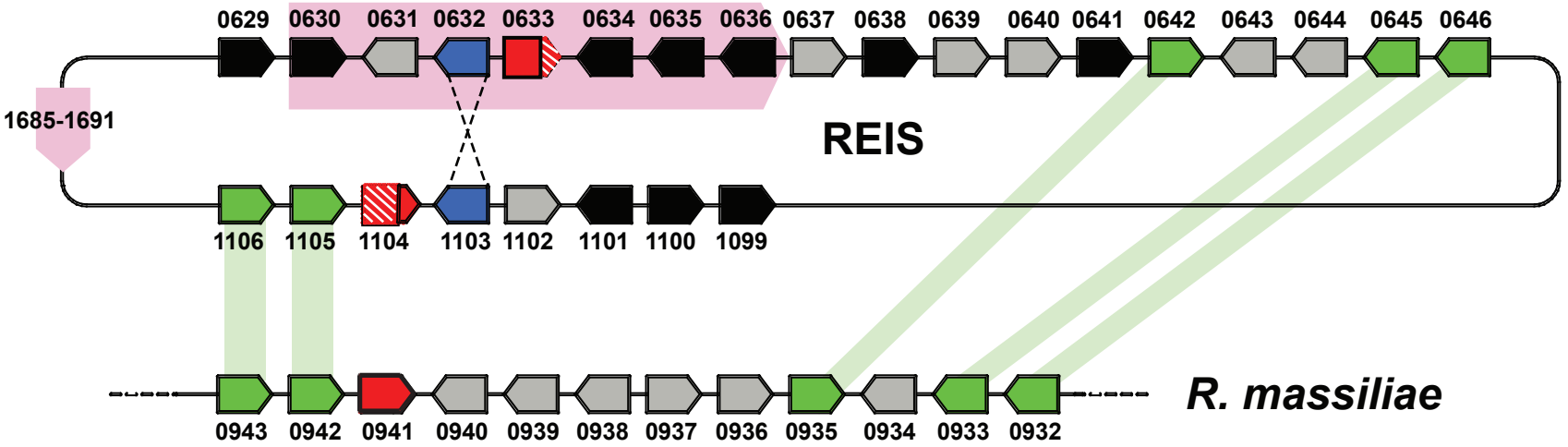

A. marginale str. Puerto Rico
A. marginale str. Florida
A. marginale str. Virginia
A. marginale str. Mississippi
A. marginale str. St. Maries
A. centrale str. Israel
A. phagocytophilum str. HZ
E. ruminantium str. Welgevonden e
E. ruminantium str. Welgevonden o
E. ruminantium str. Gardel
E. chaffeensis str. Arkansas
E. chaffeensis str. Sapulpa
E. canis str. Jake
endosymb. Drosophila melanogaster
endosymb. Drosophila willistoni
endosymb. Muscidifurax uniraptor
Wolbachia sp. wRi
endosymb. Drosophila ananassae
endosymb. Drosophila simulans
endosymb. C. quinquefasciatus JHB
endosymb. C. quinquefasciatus Pel
endosymb. str. TRS of B. malayi
N. sennetsu str. Miyayama
N. risticii str. Illinois
R. rickettsii str. Sheila Smith
R. rickettsii str. Iowa
R. peacockii str. Rustic
R. africae str. ESF 5
R. sibirica str. 246
R. conorii str. Malish 7
R. massiliae str. MTU5
R. endosymb. I. scapularis
R. akari str. Hartford
R. felis str. URRWXCal2
R. prowazekii str. Rp22
R. prowazekii str. Madrid E
R. typhi str. Wilmington
R. canadensis str. McKiel
R. bellii str. OSU 85 389
R. bellii str. RML369-C
O. tsutsugamushi str. Ikeda
O. tsutsugamushi str. Boryong
Candidatus P. ubique str. HTCC1002
Candidatus P. ubique str. HTCC1062
Candidatus Pelagibacter sp. HTCC7211
alpha proteobacterium str. HIMB114
Bradyrhizobium japonicum str. USDA 110
Pseudovibrio sp. JE062
Anaplasma
Ehrlichia
Wolbachia
Neorickettsia
Anaplasmataceae
Rickettsia
Orientia
Rickettsiaceae
SFG
TRG
TG
AG
STG
RICKETTSIALES
SAR11

Fig. S5. Comparative analysis of REIS RickA-encoding ORFs with RickA proteins from 12 *Rickettsia* genomes. (A) Genomic distribution of REIS *rickA* ORFs relative to the *rickA*-encoding region of *R. massiliae* (RMA). Filled pentagons depict genes (indicating direction of transcription). *rickA* ORFs are colored red, with missing fragments in the REIS genes hashed. RMA *rickA* (RMA_0941) is full length, whereas REIS contains three partial copies: REIS_1104 encoding the 5' end of *rickA* and two dispersed identical copies of the 3' end of the gene (REIS_0633 and REIS_1688). The location of the three REIS *rickA* fragments suggests the following scenario: the insertion sequence (IS) ISPg3 (whose transposase gene is colored blue) inserted into and split *rickA*. Subsequent homologous recombination between the inserted ISPg3 and another copy of the IS (dashed lines between REIS_0632 and REIS_1103) inverted a large portion of the REIS genome. Finally, the region of the 3' *rickA* fragment (pink shading) was duplicated into a distant region of the genome. Genes putatively from other mobile genetic elements are shown in black. Orthologous genes in the vicinity of *rickA* across both genomes are shown in green and connected by green shading. All other genes are shown in gray. (B) Schema of the RickA protein of *Rickettsia* spp, as originally defined [1, 2]. Red, G-actin binding domain; green, proline rich region; orange, WASP (Wiskott–Aldrich Syndrome protein) homology 2 region; blue, central domain; brown, acidic domain. Dashed box depicts the region of REIS RickA that was interrupted by ISPg3 insertion. (C) Protein sequence alignment of RickA proteins from 12 *Rickettsia* genomes and the three ORFs encoding RickA in the REIS genome. Coloring follows the domains illustrated in panel B, with coordinates from the complete protein alignment. Within the Pro-rich region, numbers in parentheses indicate total number of proline residues. Sequences were aligned with MUSCLE v3.6 using default parameters [3, 4].

1. Gouin, E., et al., *The RickA protein of Rickettsia conorii activates the Arp2/3 complex.* Nature, 2004. **427**(6973): p. 457-61.
2. Jeng, R.L., et al., *A Rickettsia WASP-like protein activates the Arp2/3 complex and mediates actin-based motility.* Cell Microbiol, 2004. **6**(8): p. 761-9.
3. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
4. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

A



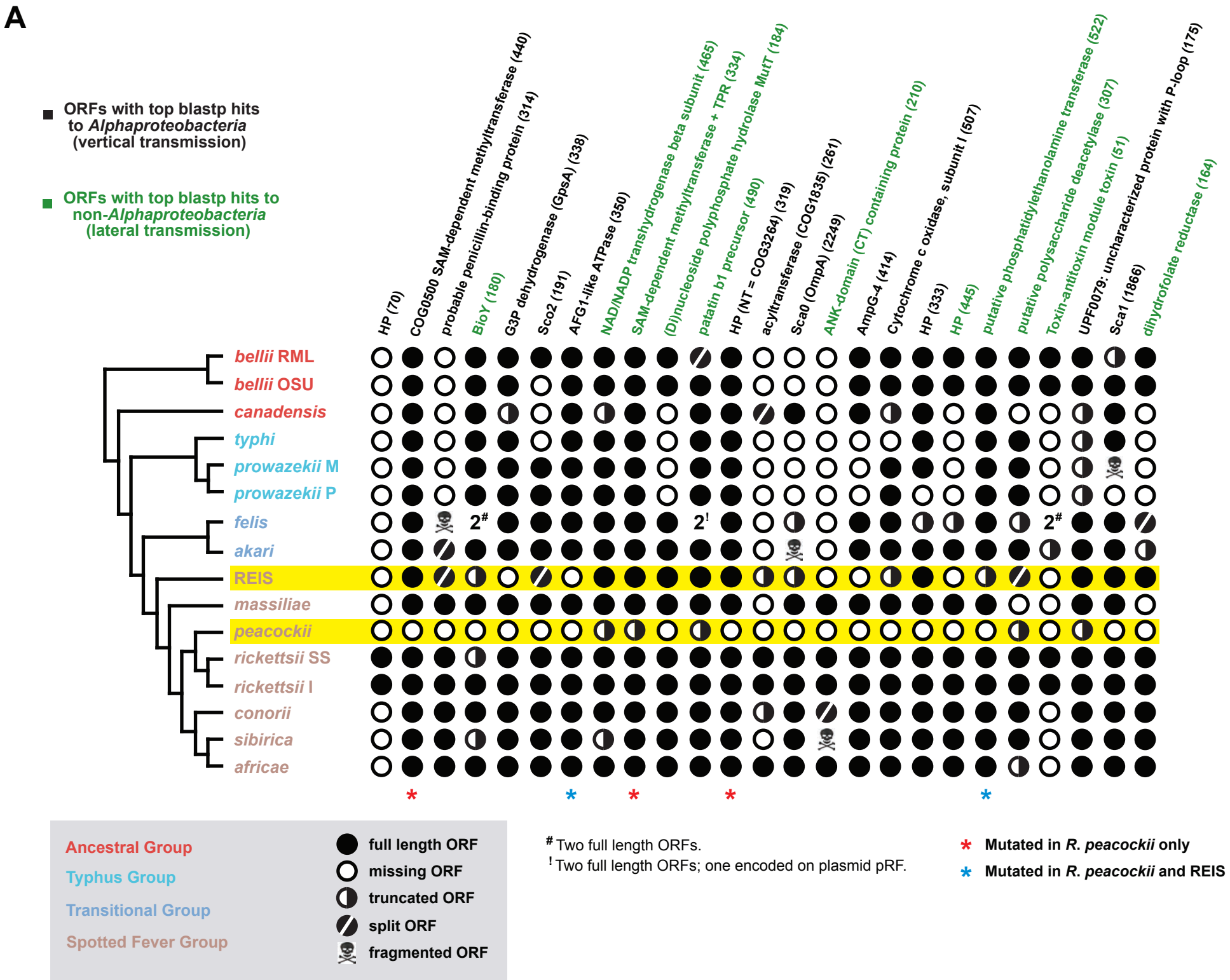
B

C

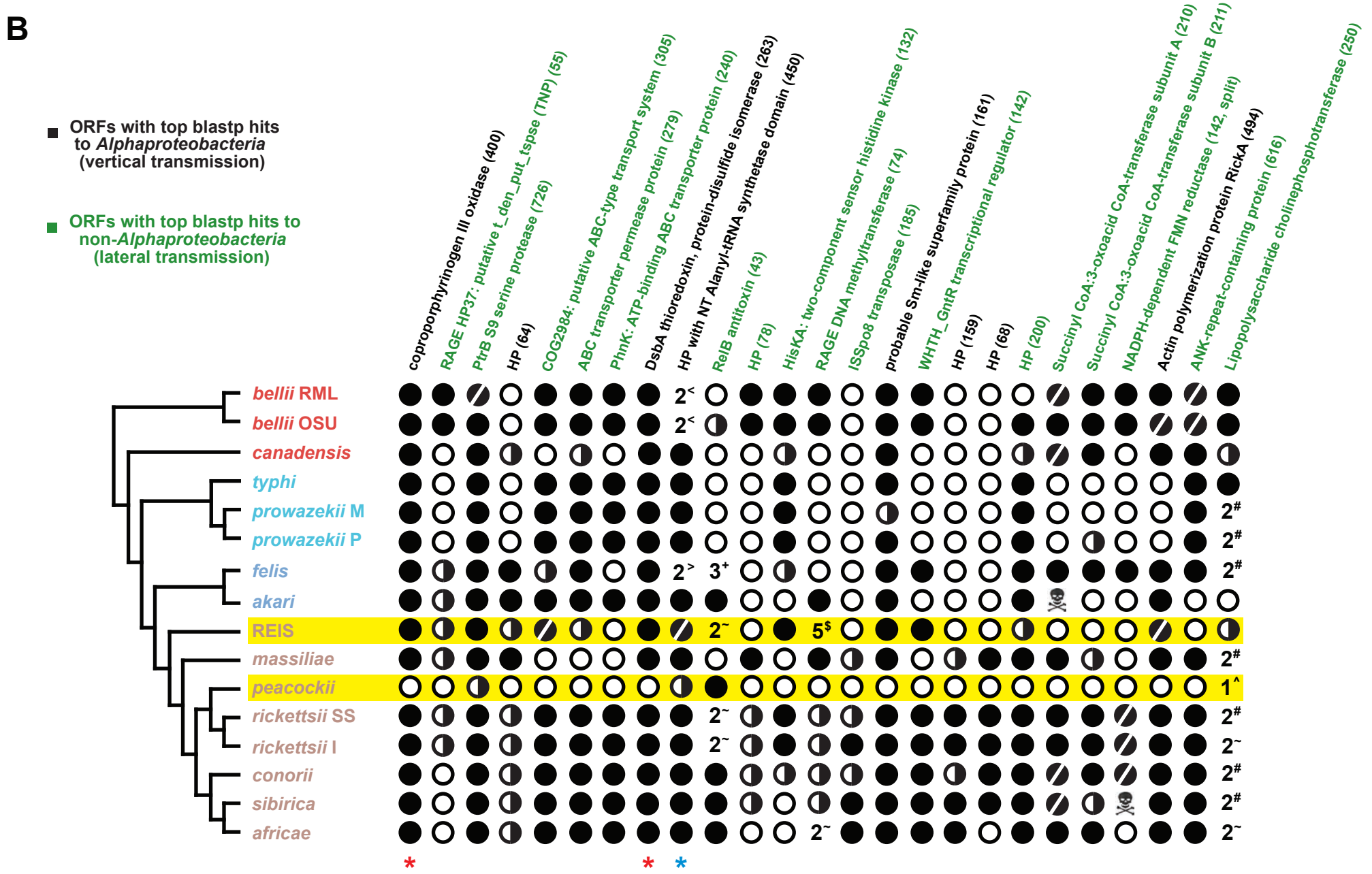
	112-127	343-426	455-500	520-537	547-556	
R. bellii R	LSVADKSGPLKQELQK	70 (27)	TTNLMKQIQGGF----------NLKKIEY	DPI--IAALNKIRSAKV	SGTDSGWASD	518
R. bellii O	----------	0 (0)	----------	----------	----------	166
R. canadensis	YNIVAKSAPLKQALQE	69 (45)	TSDLMKEIVGPRNLKEVKKIDAKAQDPRDLLLQSIRGEHKLKKVEF	NKSNEIVEILARRVAME	SDSDSGNWSD	559
R. felis	YNIAEKSAPLKQELQE	38 (20)	TSDLMREIAGPKNLRKVEKTDVKTQDSRDLLLQSIRGEHKLRKVEF	SKPNGVASILARRVAME	SESDSGNWSD	529
R. akari	YNIAAKSAPLKQELQE	35 (17)	TSDLMREIAGPNNLRKVEKTDVKIQDSRDLLLQSIRGEHKLRKVAF	NQPNGVASILARRVAME	SDSDSGNWSD	523
REIS_1104	YNIAEKSAPLKQELQE	22 (12)	TSDLMREIAGPKNLRKVEKTDVKAQDSRDLLLQSIRGEHKLNKPQF	----------	----------	415
REIS_0633	----------	0 (0)	----------KKVEF	NKLNGVASILARRVAME	SESDSGNWSD	97
REIS_1688	----------	0 (0)	----------KKVEF	NKLNGVASILARRVAME	SESDSGNWSD	97
R. massiliae	YSIAEKSAPLKQALQA	43 (24)	TSDLMREIVGPKKLRKVEKTDVKAQDSRDLLLQSIRGEHKLKKVEF	NKPNGVASILARRVAIE	SESDSGNWSD	532
R. raoultii	YNIAEKSAPLKQELQE	77 (44)	TSDLMREIAGPKKLRKVEKTDVKAQDSRDLLLQSIRGEHKLKKVEF	NKPNGIASILARRVAME	SESDSGNWSD	565
R. rickettsii S	YNIAEKSAPLKQALQE	41 (29)	TSDLMREIAGPK----------KLKKVEF	NKPSGLESIFARRVAIE	SESDSGNWSD	494
R. rickettsii I	YNIAEKSAPLKQALQE	28 (21)	TSDLMREIAGPK----------KLKKVEF	NKPSGLESIFARRVAIE	SESDSGNWSD	481
R. conorii	YNIAEKSAPLKQALQE	62 (38)	TSDLMREIAGPK----------KLKKVEF	NALSGLESIFARRAVIK	SESDSGNWSD	520
R. sibirica	YNIAEKSAPLKQALQE	71 (43)	TSDLMREIAGPK----------KLKKVEF	NKPSGLESIFARRAAIE	SESDSGNWSD	526
R. africae	YNIAEKFAPLKQALQE	42 (30)	TSALMREIAGPK----------KLKKVEF	NKPSGLESILARRIAIE	SESDSGNWSD	497
	* **** **		* ** * * * * * * ********** * *	*	* *** **	

Fig. S6. Characteristics of 50 problematic genes of *R. peacockii* str. Rustic and comparison with homologous ORFs in 15 other *Rickettsia* genomes. (A) Genes 1-25. (B) Genes 26-50. The genes were selected from Felsheim et al. [1]. Gene product annotations are listed at the top in each panel, with black illustrating proteins with top blastp hits to other *Alphaproteobacteria* sequences (vertical transmission), and green denoting top blastp hits to non-*Alphaproteobacteria* (lateral transmission). Length of query proteins is given (aa) in parentheses. Query sequences (protein IDs) from *R. rickettsii* genomes as follows, numbering 1-50: 1, A1G_03990; 2, A1G_03950; 3, A1G_01175; 4, Rrlowa_0811; 5, A1G_03470; 6, A1G_00265; 7, A1G_01615; 8, A1G_00635; 9, A1G_00720; 10, A1G_06270; 11, A1G_05085; 12, A1G_04725; 13, A1G_07015; 14, A1G_06990; 15, A1G_04305; 16, A1G_03035; 17, A1G_02985; 18, A1G_02790; 19, A1G_02605; 20, A1G_02570; 21, A1G_00085; 22, A1G_00090; 23, A1G_00095; 24, A1G_00130; 25, A1G_00215; 26, A1G_01245; 27, A1G_01880; 28, A1G_02165; 29, A1G_02530; 30, A1G_02820; 31, A1G_02825; 32, A1G_02830; 33, A1G_03355; 34, A1G_03530; 35, A1G_04035; 36, A1G_04170; 37, A1G_04290; 38, A1G_04355; 39, A1G_04365; 40, A1G_04605; 41, A1G_04620; 42, A1G_04660; 43, A1G_04775; 44, A1G_04970; 45, A1G_04995; 46, A1G_05000; 47, A1G_05005/A1G_05010; 48, A1G_05015; 49, A1G_05165; 50, A1G_05855. All other information is provided at the bottom of each panel.

1. Felsheim, R.F., T.J. Kurtti, and U.G. Munderloh, *Genome sequence of the endosymbiont Rickettsia peacockii and comparison with virulent Rickettsia rickettsii: identification of virulence factors.* PLoS One, 2009. **4**(12): p. e8361.


A
ORFs with top blastp hits to Alphaproteobacteria (vertical transmission)
ORFs with top blastp hits to non-Alphaproteobacteria (lateral transmission)
HP (70)
COG0500 SAM-dependent methyltransferase (440)
probable penicillin-binding protein (314)
BioY (180)
G3P dehydrogenase (GpsA) (338)
Sco2 (191)
AFG1-like ATPase (350)
NAD/NADP transhydrogenase beta subunit (465)
SAM-dependent methyltransferase + TPR (334)
(Di)nucleoside polyphosphate hydrolase MutT (184)
patatin b1 precursor (490)
HP (NT = COG3264) (319)
acyltransferase (COG1835) (261)
Sca0 (OmpA) (2249)
ANK-domain (CT) containing protein (210)
AmpG-4 (414)
Cytochrome c oxidase, subunit I (507)
HP (333)
HP (445)
putative phosphatidylethanolamine transferase (522)
putative polysaccharide deacetylase (307)
Toxin-antitoxin module toxin (51)
UPF0079: uncharacterized protein with P-loop (175)
Sca1 (1866)
dihydrofolate reductase (164)
bellii RML
bellii OSU
canadensis
typhi
prowazekii M
prowazekii P
felis
akari
REIS
massiliae
peacockii
rickettsii SS
rickettsii I
conorii
sibirica
africae
2#
2!
2#
Ancestral Group
Typhus Group
Transitional Group
Spotted Fever Group
full length ORF
missing ORF
truncated ORF
split ORF
fragmented ORF
Two full length ORFs.
! Two full length ORFs; one encoded on plasmid pRF.
Mutated in R. peacockii only
Mutated in R. peacockii and REIS

B



Ancestral Group
Typhus Group
Transitional Group
Spotted Fever Group

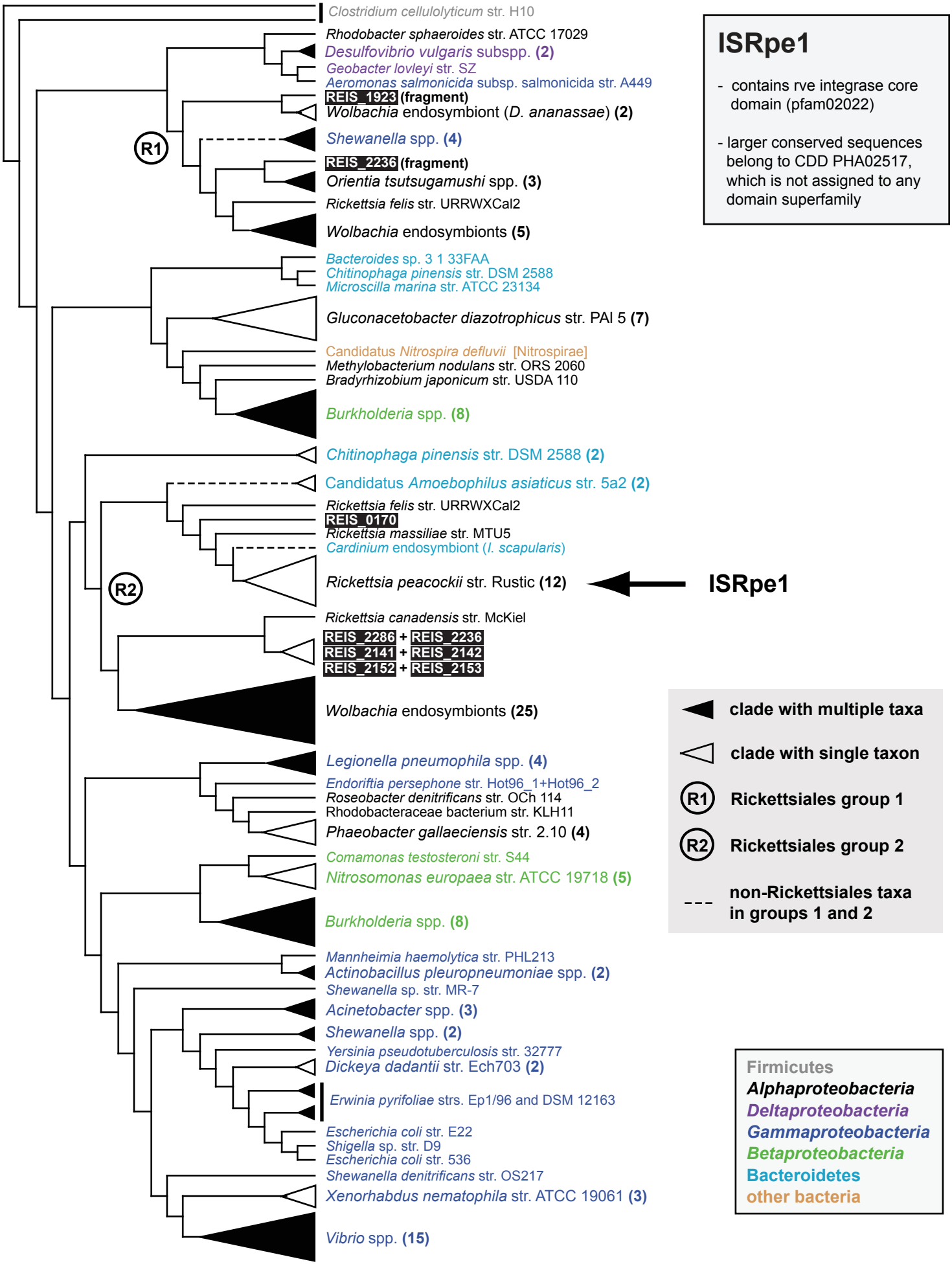
- full length ORF
- missing ORF
- truncated ORF
- split ORF
- fragmented ORF

Two full length ORFs.
! Two full length ORFs; one encoded on plasmid pRF.
< Two full length ORFs (one conserved, one divergent).
> One full length ORF (divergent), one truncated.
+ Three full length ORFs.
~ One full length ORF, one truncated.
$ Four full length ORFs, one truncated, additional pseudogenes.
^ Two ORFs recombined into one.

* Mutated in *R. peacockii* only
* Mutated in *R. peacockii* and REIS

Fig. S7. Compilation and phylogeny estimation of 216 ISRPe1 and ISRpe1-like transposase sequences. The integrase core domain of these sequences belongs to the rve superfamily (pfam02022). All sequences share homology outside of the integrase domain, and collectively are grouped in the conserved domain PHA02517, which is not assigned to any domain superfamily. A total of 65 putative ISRpre1 sequences were retrieved using *R. peacockii* ISRpe1 as a query against the Rickettsiales database (taxid:766). Split ORFs were merged resulting in a total of 58 ISRpe1 sequences. These Rickettsiales sequences were combined with 158 blastp subjects acquired using the same query sequence against the NCBI non-redundant protein database excluding taxid:766. The total 216 ISRpe1 and ISRpe1-like sequences were aligned with MUSCLE v3.6 using default parameters [1, 2]. A phylogeny was estimated in PAUP* v4.0b10 (Altivec) under parsimony [3]. A heuristic search was implemented employing 250 random sequence additions, saving 10 trees per replication. A majority rule consensus tree was constructed for 170 equally parsimonious trees of tree score 4458. All nodes shown on the tree were present in all 170 trees, with collapsed clades (open and filled triangles) containing nodes not present in all trees.

1. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
2. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
3. Swofford, D., *PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4 ed.* , 1999, Sinauer: Sunderland, MA.


Clostridium cellulolyticum str. H10
Rhodobacter sphaeroides str. ATCC 17029
Desulfovibrio vulgaris subspp. (2)
Geobacter lovleyi str. SZ
Aeromonas salmonicida subsp. salmonicida str. A449
REIS_1923 (fragment)
Wolbachia endosymbiont (D. ananassae) (2)
R1
Shewanella spp. (4)
REIS_2236 (fragment)
Orientia tsutsugamushi spp. (3)
Rickettsia felis str. URRWXCal2
Wolbachia endosymbionts (5)
Bacteroides sp. 3 1 33FAA
Chitinophaga pinensis str. DSM 2588
Microscilla marina str. ATCC 23134
Gluconacetobacter diazotrophicus str. PAl 5 (7)
Candidatus Nitrospira defluvii [Nitrospirae]
Methylobacterium nodulans str. ORS 2060
Bradyrhizobium japonicum str. USDA 110
Burkholderia spp. (8)
Chitinophaga pinensis str. DSM 2588 (2)
Candidatus Amoebophilus asiaticus str. 5a2 (2)
Rickettsia felis str. URRWXCal2
REIS_0170
Rickettsia massiliae str. MTU5
Cardinium endosymbiont (I. scapularis)
R2
Rickettsia peacockii str. Rustic (12)
ISRpe1
Rickettsia canadensis str. McKiel
REIS_2286 + REIS_2236
REIS_2141 + REIS_2142
REIS_2152 + REIS_2153
Wolbachia endosymbionts (25)
Legionella pneumophila spp. (4)
Endoriftia persephone str. Hot96_1+Hot96_2
Roseobacter denitrificans str. OCh 114
Rhodobacteraceae bacterium str. KLH11
Phaeobacter gallaeciensis str. 2.10 (4)
Comamonas testosteroni str. S44
Nitrosomonas europaea str. ATCC 19718 (5)
Burkholderia spp. (8)
Mannheimia haemolytica str. PHL213
Actinobacillus pleuropneumoniae spp. (2)
Shewanella sp. str. MR-7
Acinetobacter spp. (3)
Shewanella spp. (2)
Yersinia pseudotuberculosis str. 32777
Dickeya dadantii str. Ech703 (2)
Erwinia pyrifoliae strs. Ep1/96 and DSM 12163
Escherichia coli str. E22
Shigella sp. str. D9
Escherichia coli str. 536
Shewanella denitrificans str. OS217
Xenorhabdus nematophila str. ATCC 19061 (3)
Vibrio spp. (15)
ISRpe1
- contains rve integrase core domain (pfam02022)
- larger conserved sequences belong to CDD PHA02517, which is not assigned to any domain superfamily
clade with multiple taxa
clade with single taxon
R1 Rickettsiales group 1
R2 Rickettsiales group 2
non-Rickettsiales taxa in groups 1 and 2
Firmicutes
Alphaproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Betaproteobacteria
Bacteroidetes
other bacteria