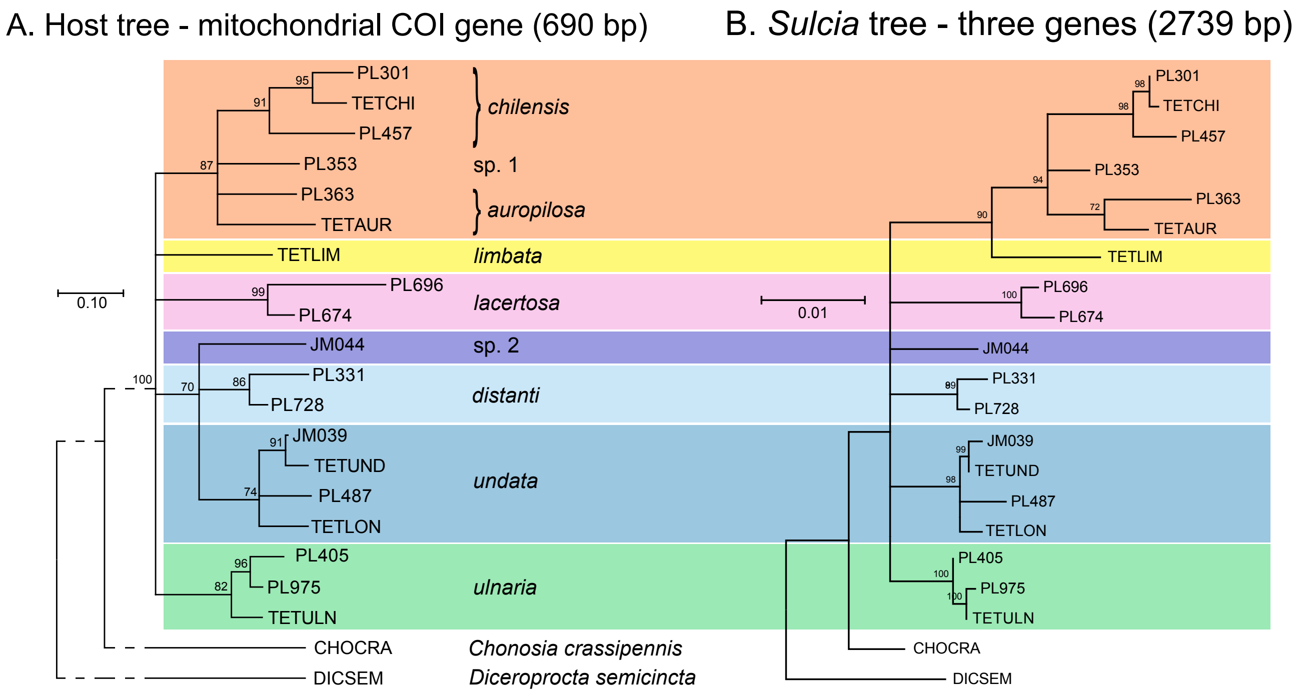
Multiple origins of interdependent endosymbiotic complexes in a genus of cicadas

Piotr Łukasik, Katherine Nazario, James T. Van Leuven, Matthew A. Campbell, Mariah Meyer, Anna Michalik, Pablo Pessacq, Chris Simon, Claudio Veloso, John P. McCutcheon

SI Appendix



A. Host tree - mitochondrial COI gene (690 bp)

B. *Sulcia* tree - three genes (2739 bp)

Figure S1. Maximum likelihood phylogenies of cicadas from the genus *Tettigades* (A) and of their *Sulcia* symbionts (B), based on partial sequences of mitochondrial cytochrome oxidase I (*COI*) gene, or on concatenated partial sequences of three protein-coding genes, *rpoB*, *rplB*, and *bamA*. The phylogenies have been constrained based on multi-gene phylogenies for samples with genomes sequenced (see Fig. S2). Bootstrap values of 70% or more are shown as percentages above the nodes, and nodes with lower support have been collapsed. Color ranges represent species or species groups in which the ancestral state was a single lineage of *Hodgkinia*.

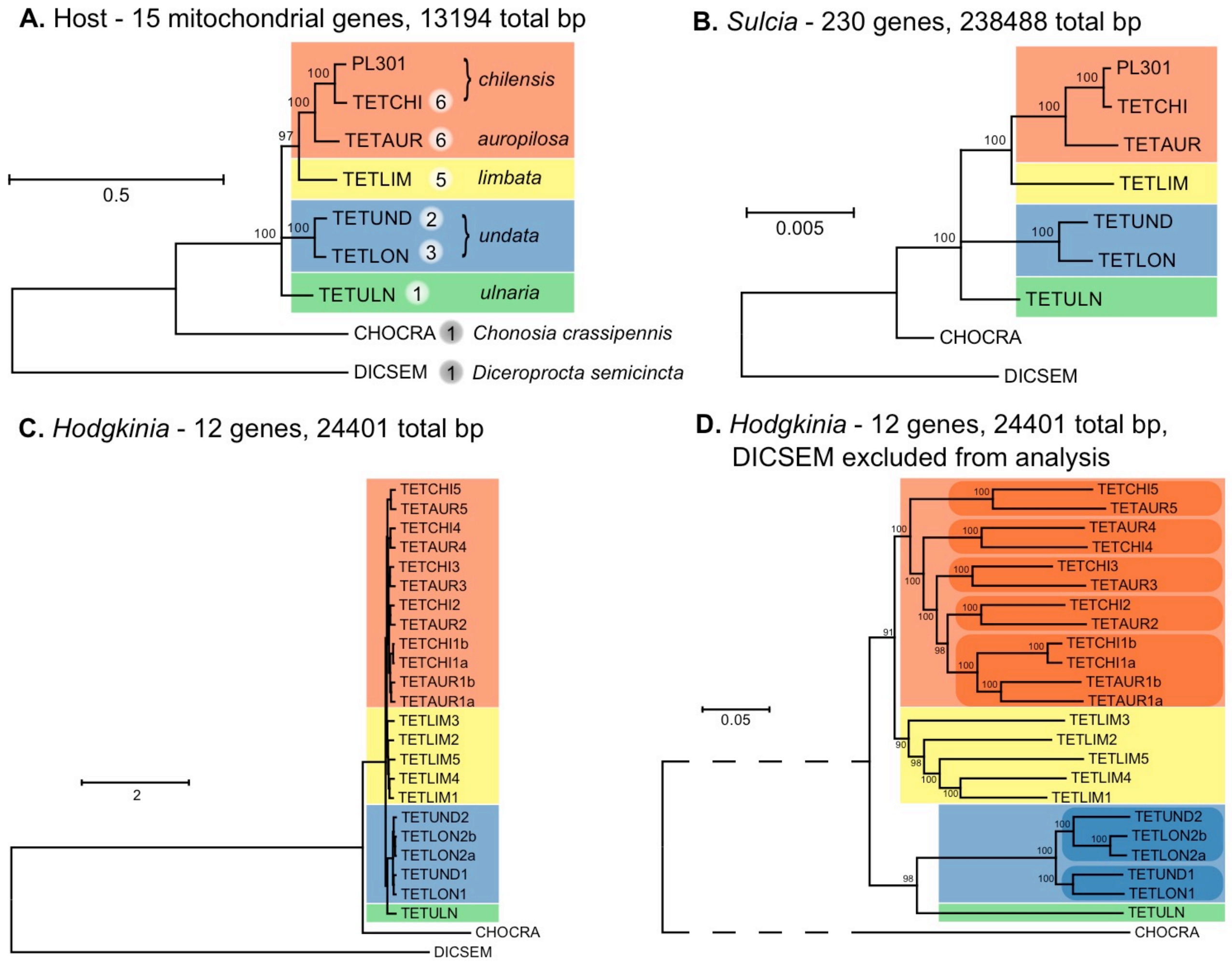
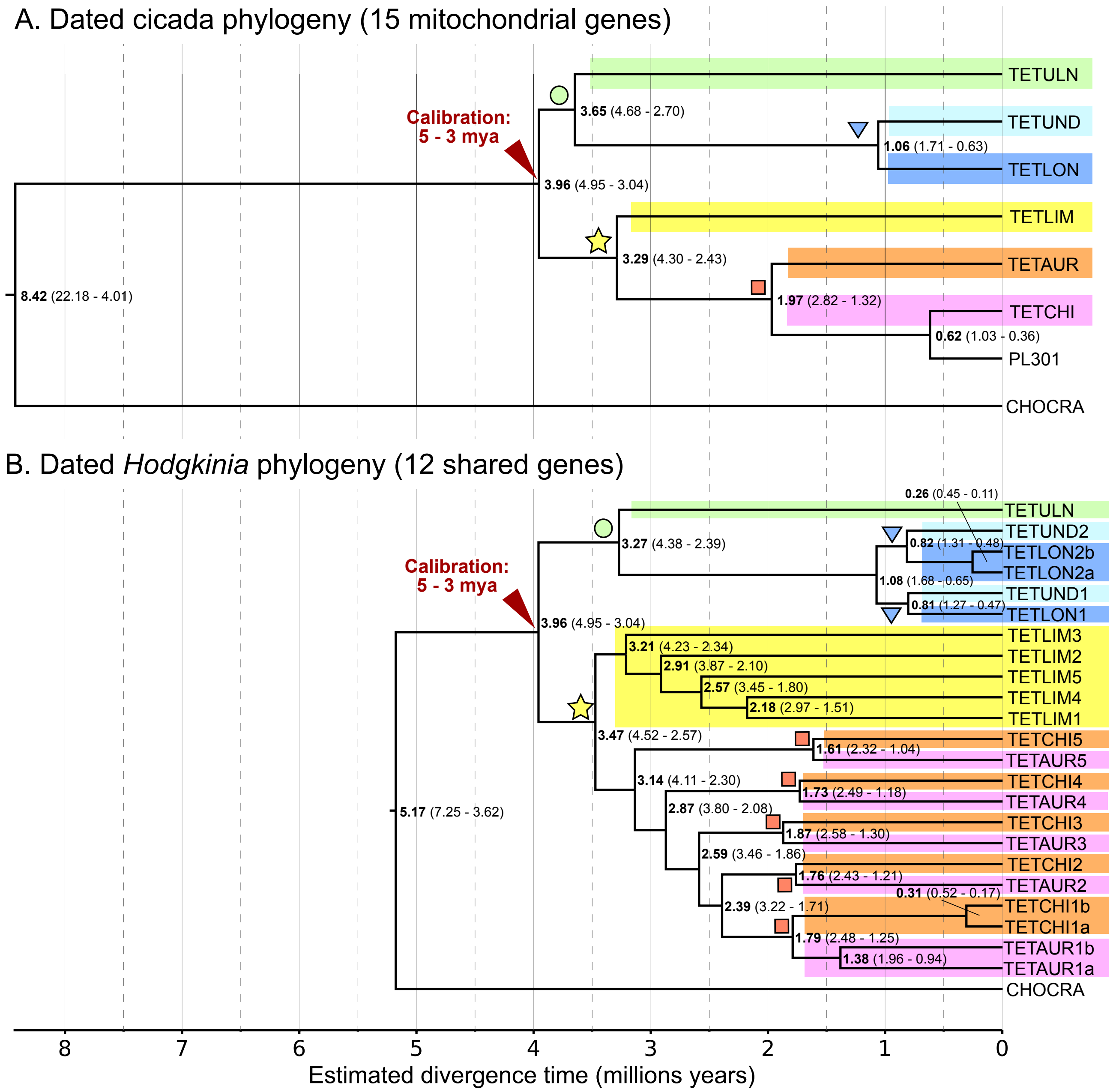


Figure S2. Maximum likelihood phylogenies of cicadas with sequenced bacteriome metagenomes, and of their *Sulcia* and *Hodgkinia* symbionts. Phylogenies are based on all protein-coding, rRNA and ncRNA genes shared across their genomes. Nodes with bootstrap support below 90% were collapsed, and support values of 90% or greater are shown as the percentage above nodes. Color ranges represent species groups in which the ancestral state was a single lineage of *Hodgkinia*, and correspond to those in Fig. 2. In panel A, numbers after leaf labels represent the number of distinct *Hodgkinia* genomes present in the studied cicada. Trees in panel C and D were estimated using the same data, with the exception of the divergent outgroup (DICSEM) that was included in the former dataset but excluded from the latter. Colored ovals in panel D indicate *Hodgkinia* clades that include at least one strain from each host specimen in a given species group, and which were thus present in the last common ancestor of that species group. For *T. chilensis* specimen PL301, we assembled mitogenome and *Sulcia* genome but not *Hodgkinia* genomes, and thus this sample is absent from panels C and D.



A. Dated cicada phylogeny (15 mitochondrial genes)

B. Dated *Hodgkinia* phylogeny (12 shared genes)

Figure S3. The estimates of the timing of divergence of the studied cicada populations, and of the splits of their *Hodgkinia* symbionts, as calculated by PhyloBayes. A single calibration point was used (indicated). Ranges near the nodes represent 95% confidence intervals for the times of respective nodes (in million years). Colored symbols near the nodes represent the cicada divergence events, as opposed to *Hodgkinia* splits within the host; because the last common ancestor of TETUND-TETLON and of TETCHI-TETAUR hosted more than one *Hodgkinia* lineage, the corresponding host divergence events are replicated on the symbiont phylogeny.

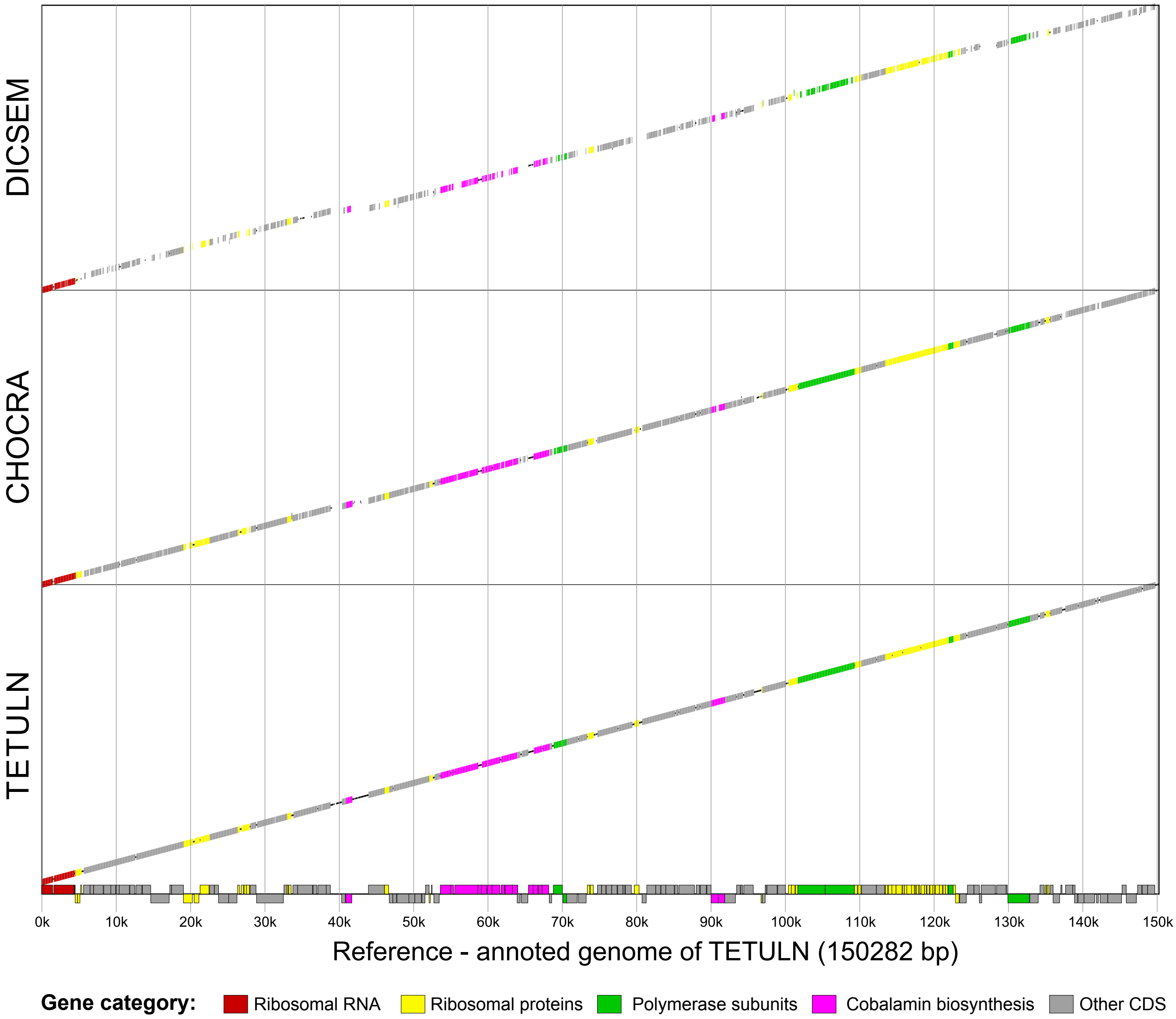


Figure S4. Alignments of Hodgkinia genomes from three cicada species that host a single symbiont lineage against the genome of the Tettigades ulnaria (TETULN) symbiont, conducted in the protein space using tblastx and custom scripts. Genes identified as functional are mapped onto alignments. The lower panel represents the alignment of the TETULN genome against itself, and can serve as a reference for figures S5-S9. The alignment of the DICSEM genome against TETULN is somewhat fragmented because of low similarity at the protein level between the two genomes.

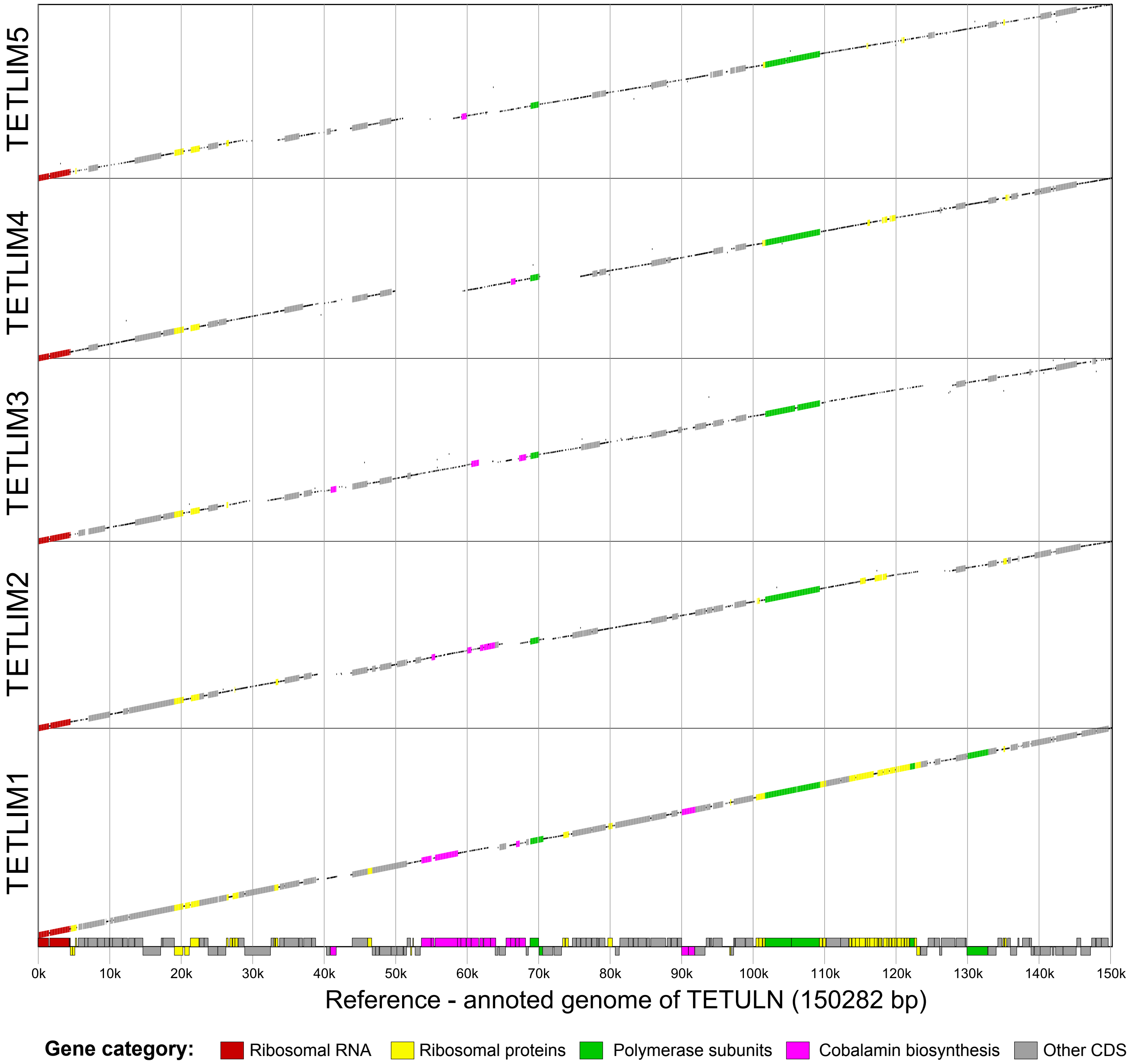


Figure S5. Alignments of *Hodgkinia* genomes from *Tettigades limbata* (TETLIM) against the single *Hodgkinia* genome from *T. ulnaria* (TETULN), conducted in the protein space using tblastx and custom scripts. Genes identified as functional are mapped onto alignments.

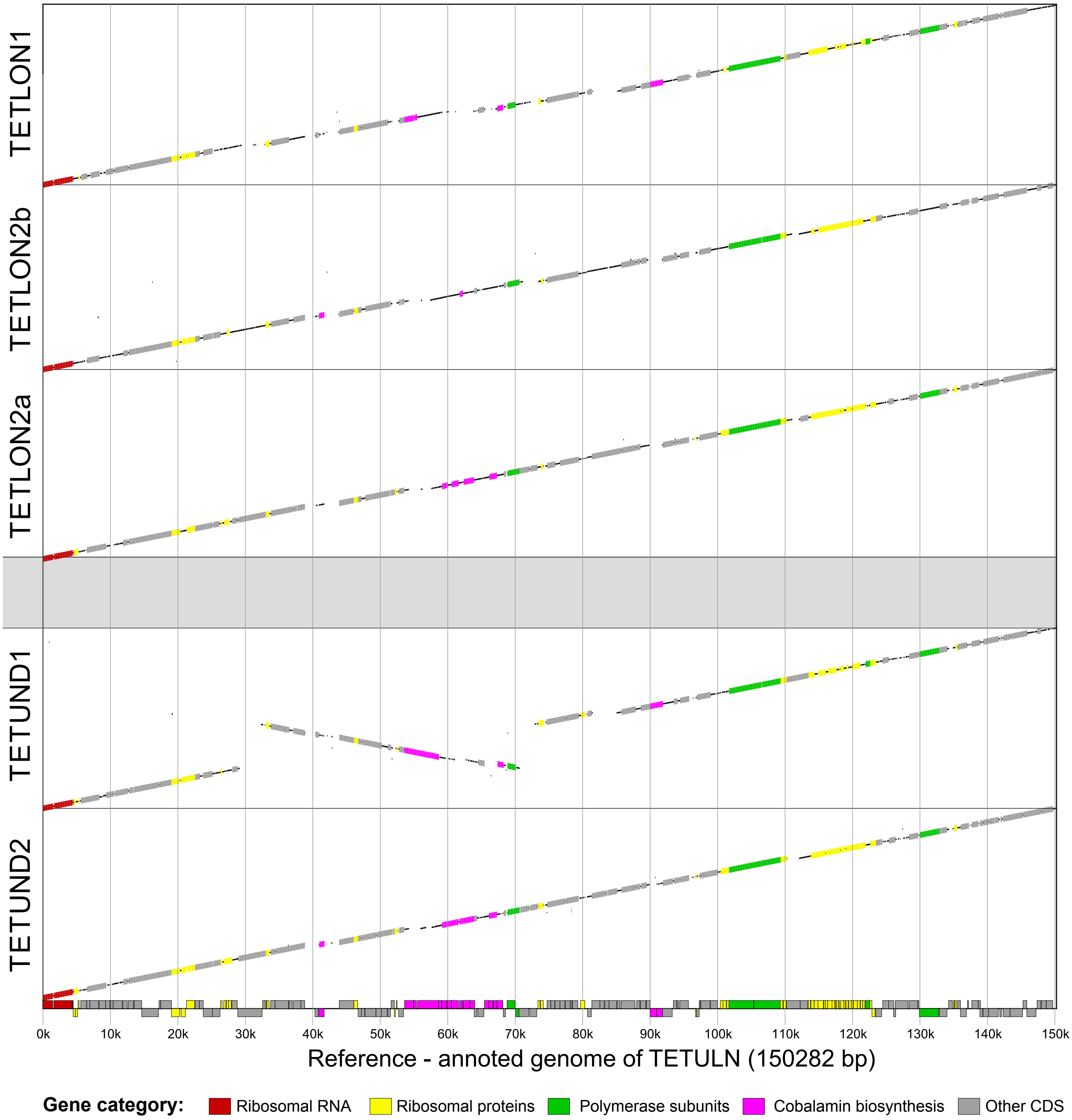
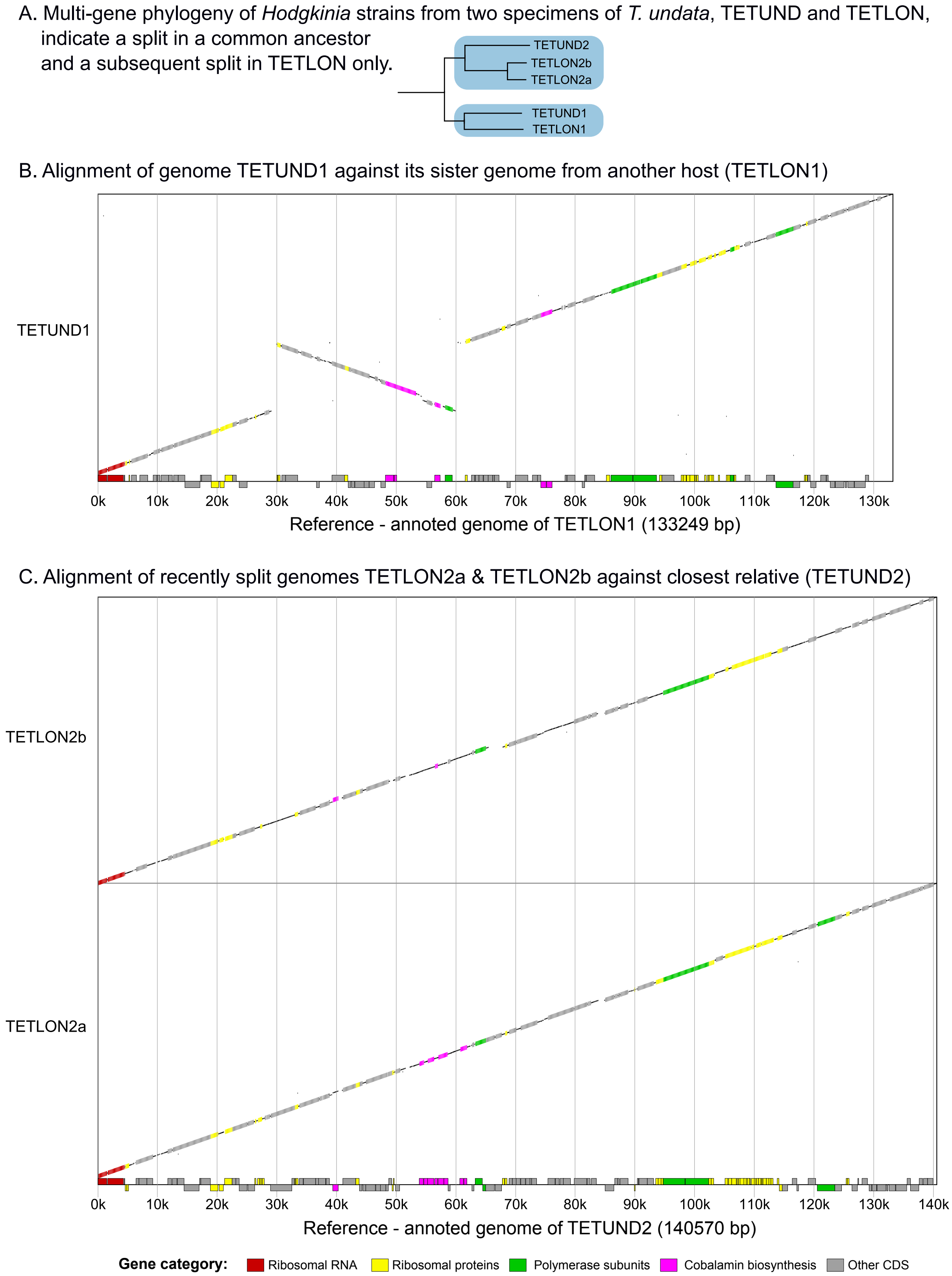


Figure S6. Alignments of *Hodgkinia* genomes from *Tettigades undata* (specimens TETUND and TETLON) against the single *Hodgkinia* genome from *T. ulnaria* (TETULN), conducted in the protein space using tblastx and custom scripts. Genes identified as functional are mapped onto alignments.



A. Multi-gene phylogeny of *Hodgkinia* strains from two specimens of *T. undata*, TETUND and TETLON, indicate a split in a common ancestor and a subsequent split in TETLON only.

B. Alignment of genome TETUND1 against its sister genome from another host (TETLON1)

C. Alignment of recently split genomes TETLON2a & TETLON2b against closest relative (TETUND2)

Figure S7. Phylogenetic relationships among *Hodgkinia* genomes from *Tettigades undata* (specimens TETUND and TETLON), and alignments of genome(s) from each clade against the sister *Hodgkinia* genome from another specimen. Alignments were conducted in the protein space, using tblastx and custom scripts. Genes identified as functional are mapped onto alignments.

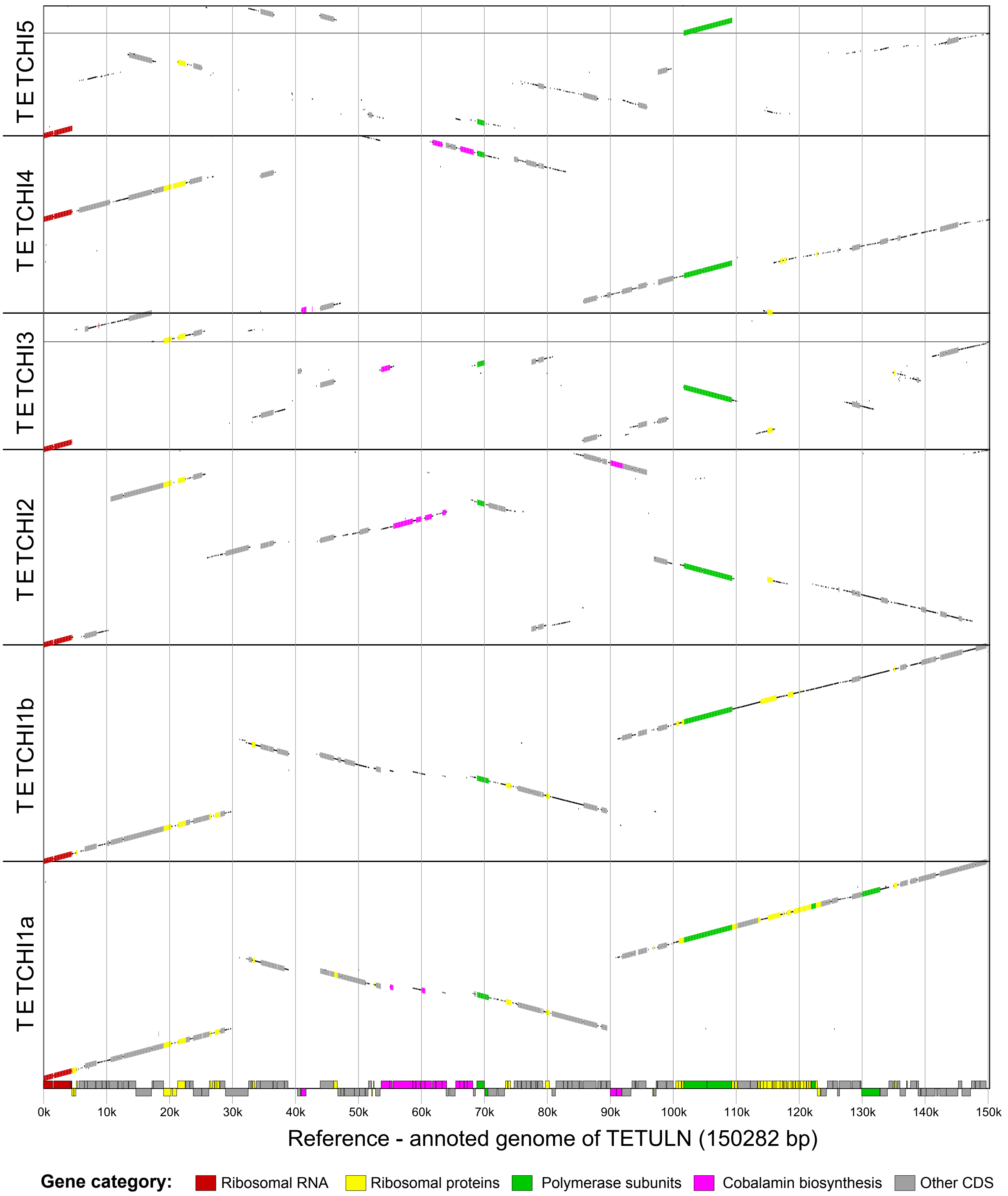


Figure S8. Alignments of six *Hodgkinia* genomes from *Tettigades chilensis* (TETCHI) against the single *Hodgkinia* genome from *T. ulnaria* (TETULN), conducted in the protein space using tblastx and custom scripts.. Genes identified as functional are mapped onto alignments. Two of the genomes, TETCHI3 and TETCHI5, consist of distinct genomic circles, or chromosomes; these chromosomes are separated with horizontal lines.

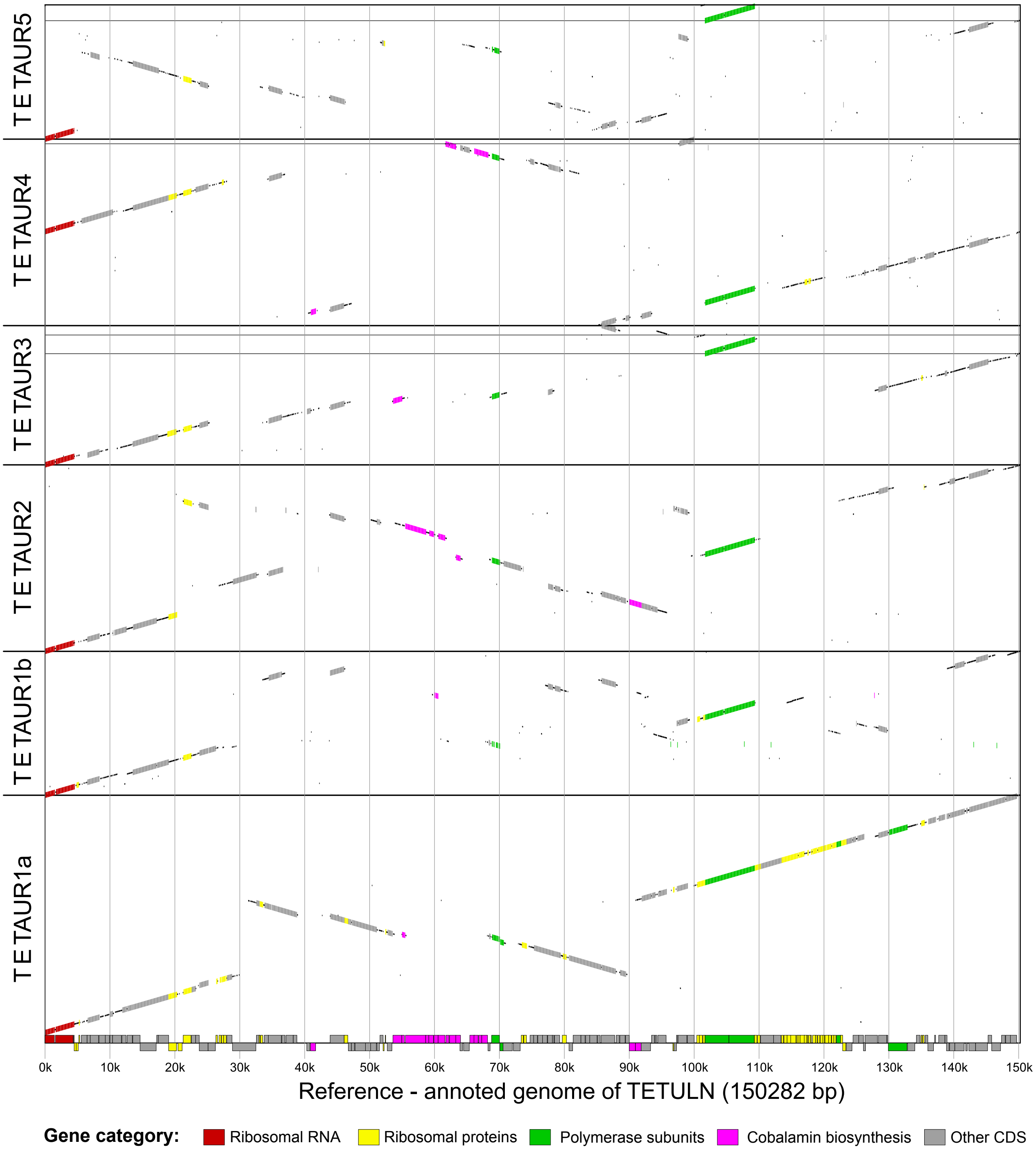
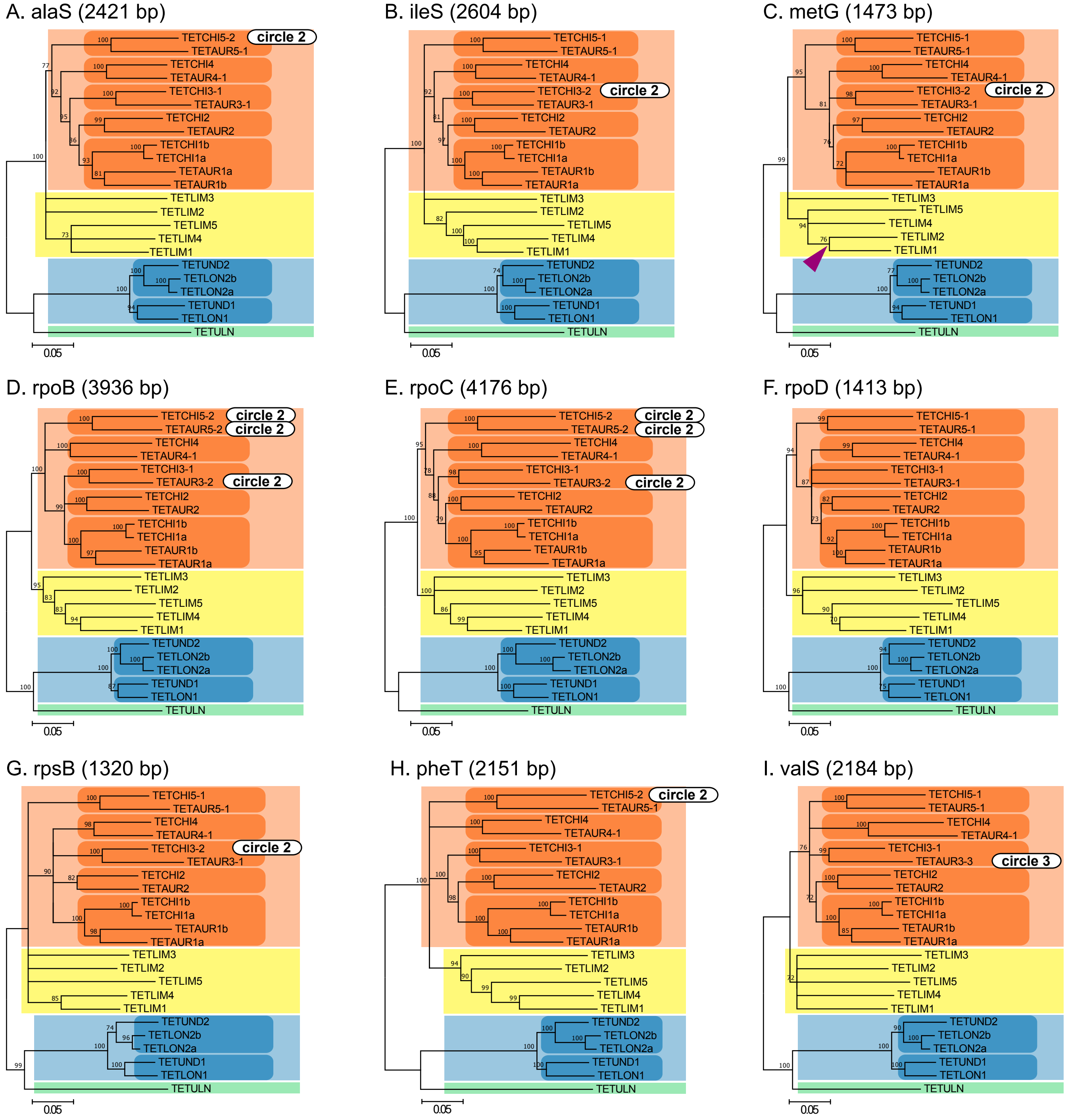


Figure S9. Alignments of six *Hodgkinia* genomes from *Tettigades auropilosa* (TETAUR) against the single *Hodgkinia* genome from *T. ulnaria* (TETULN), conducted in a protein space using tblastx and custom scripts. Genes identified as functional are mapped onto alignments. Three of the genomes (TETAUR3-5), consist of two or three distinct chromosomes each; these chromosomes are separated by horizontal lines.



A. alaS (2421 bp)

B. ileS (2604 bp)

C. metG (1473 bp)

D. rpoB (3936 bp)

E. rpoC (4176 bp)

F. rpoD (1413 bp)

G. rpsB (1320 bp)

H. pheT (2151 bp)

I. valS (2184 bp)

Figure S10. Phylogenies of protein-coding genes retained in all *Tettigades Hodgkinia* genomes. Nodes with bootstrap support below 70% were collapsed, and support values of 70% or greater are shown as the percentage above nodes. Gene trees were rooted based on the multi-gene phylogeny (see Fig. S2). Color ranges represent species groups in which the ancestral state was a single lineage of *Hodgkinia*, and correspond to those in Fig. 2. Colored ovals indicate *Hodgkinia* clades that include at least one strain from both host specimens in a given species group, and which thus must have been present in the last common ancestor of that species group. For *T. chilensis* and *T. auropilosa* genomes that consist of more than one circle, genes located on smaller circles are indicated; their phylogenetic placement is one of the arguments that these distinct circles indeed belong to a single genome. The arrowhead in panel C indicates the only node that conflicts with the multi-gene phylogeny (Fig. S2).

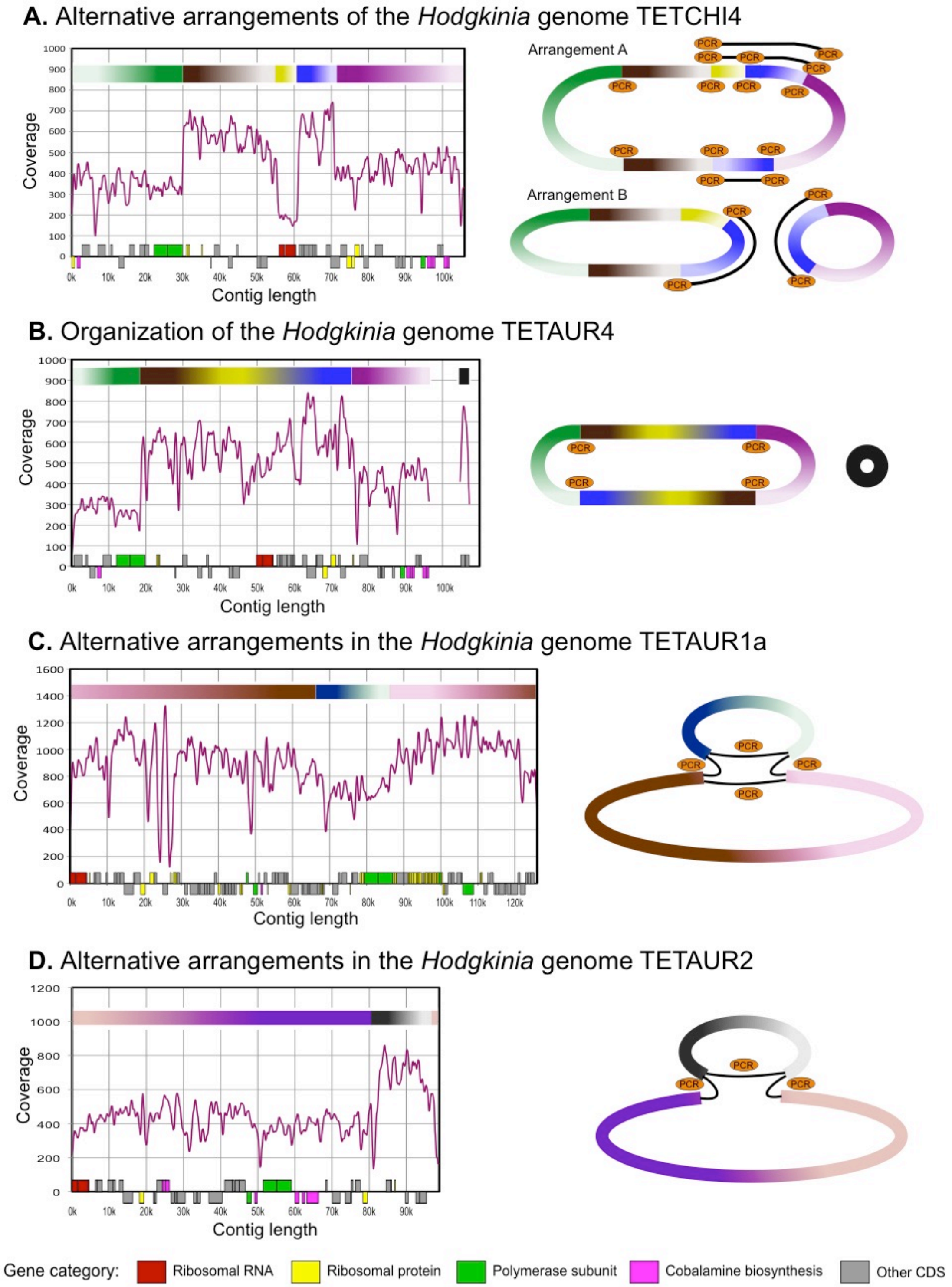


Figure S11. Verified unusual or alternative arrangements in selected *Hodgkinia* genomes. For each genome, we present coverage along the length of the annotated contig, and use color gradients to indicate different sections of the genomes and their orientation. We also indicate how these sections are connected. In all cases, we verified the section junction regions by PCR and Sanger sequencing. For genome TETCHI4, we used long-range PCR to amplify across any sections shorter than 20 kb, and after verifying product length through gel electrophoresis, we Sanger-sequenced the ends (corresponding to the junction regions).

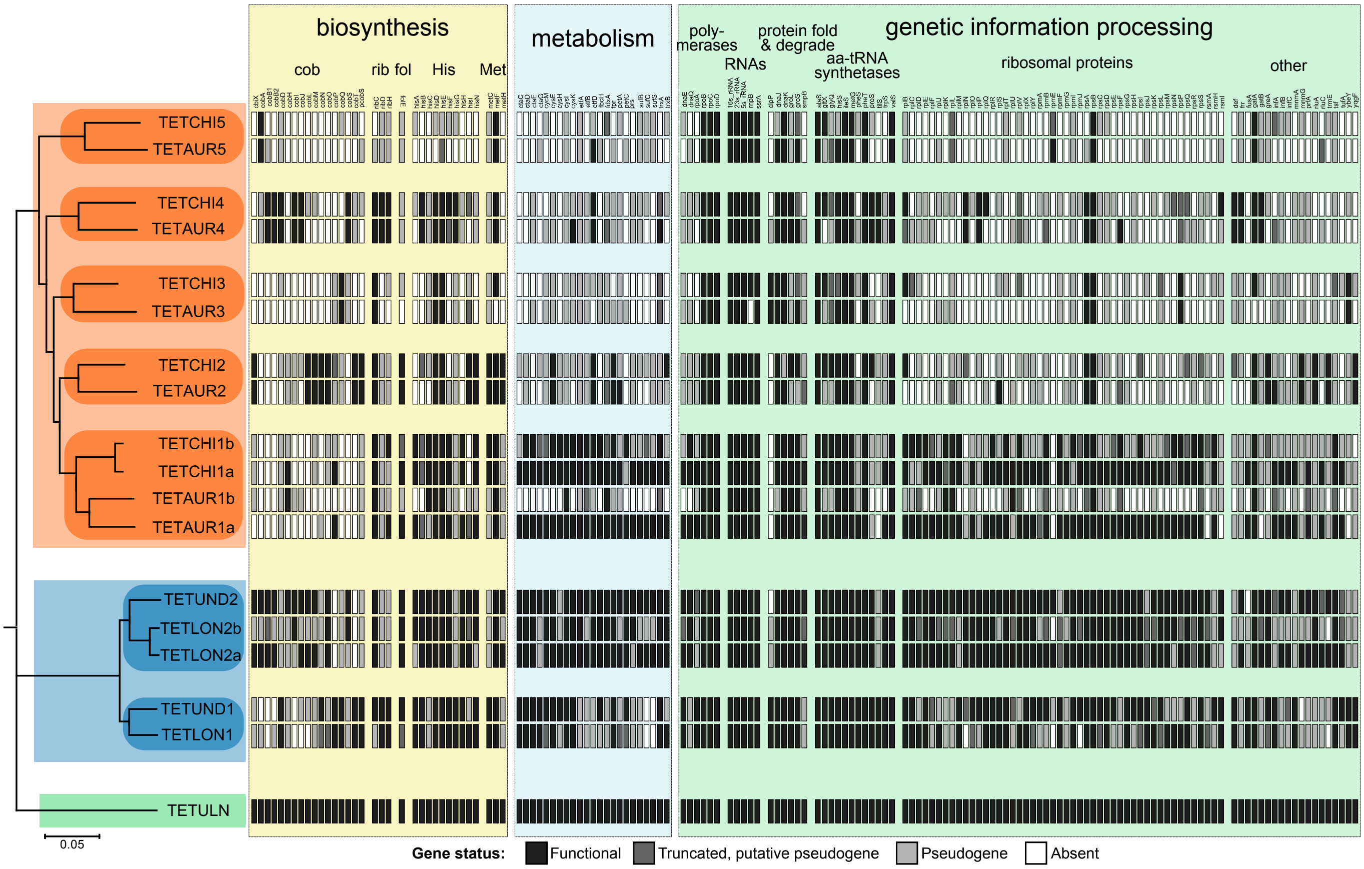


Figure S12. The list of all non-hypothetical protein-coding and RNA (other than tRNA) genes identified in individual *Hodgkinia* genomes from five cicada species. Symbiont phylogeny is based on 12 genes present in all *Hodgkinia* genomes in the current study (see Fig. S2). *Hodgkinia* has undergone four splits in the common ancestors of TETCHI and TETAUR, resulting in five clades that include at least one lineage from each of these two species (indicated with orange ovals); subsequently, there was an additional split in the ancestor of each of the studied specimens. Similarly, *Hodgkinia* has split in the ancestor of TETUND and TETLON, resulting in two clades indicated with blue ovals; subsequently, there was an additional split in the ancestor of TETLON. As in Fig. 3, the functionality classification of each gene is based on the length of the open reading frame relative to that in the genome of TETULN *Hodgkinia*.

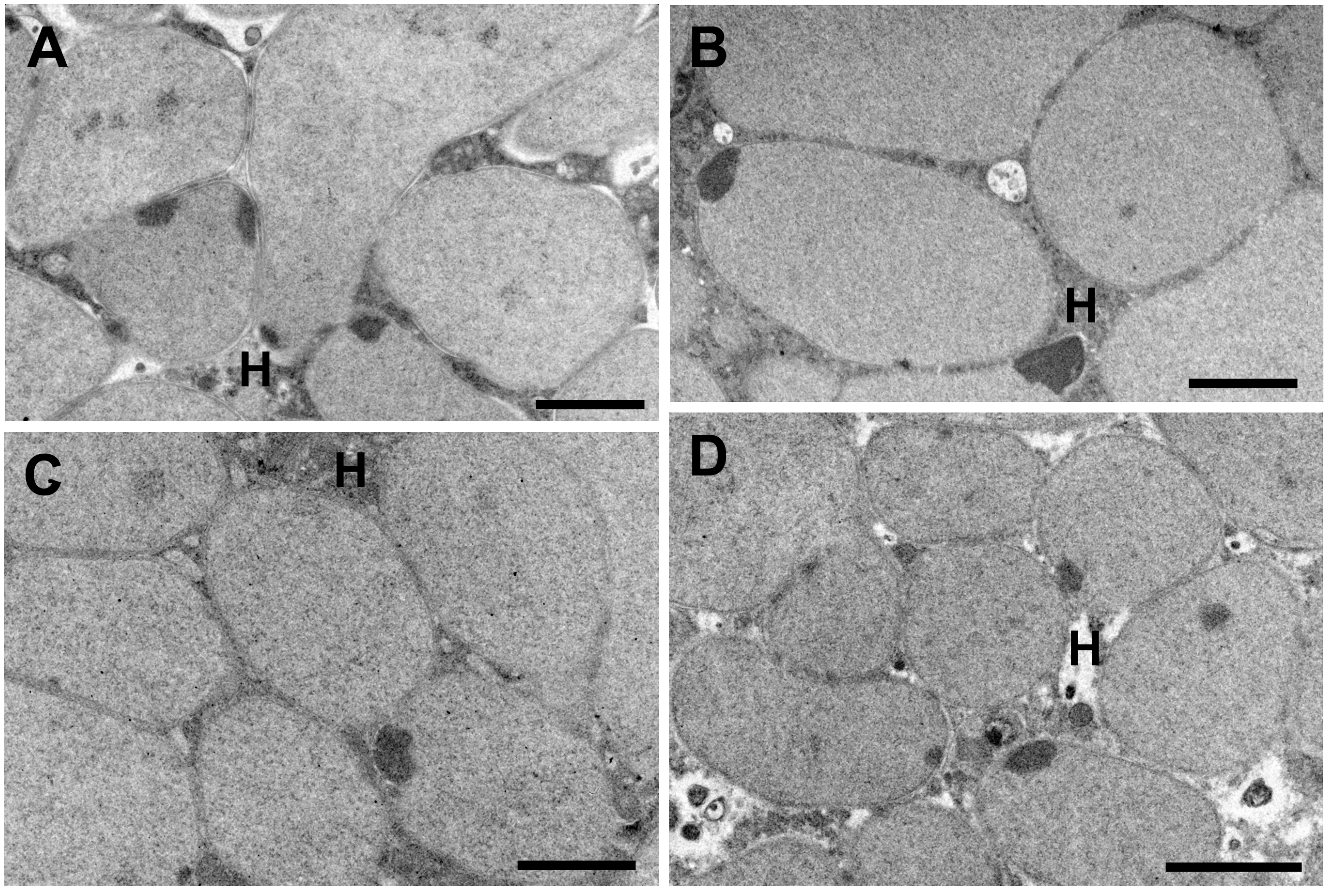


Figure S13. Similarity in morphology of *Hodgkinia* cells in four cicada species that represent distinct clades where *Hodgkinia* has undergone independent splits, and that host different numbers of *Hodgkinia* lineages. A. *Tettigades ulnaria* (1 *Hodgkinia* lineage); B. *T. undata* (2 lineages); C. *Tettigades* sp. 2 (3 lineages); D. *T. lacertosa* (4 lineages). H - *Hodgkinia*-filled syncytium; TEM, scale bar - 2µm.

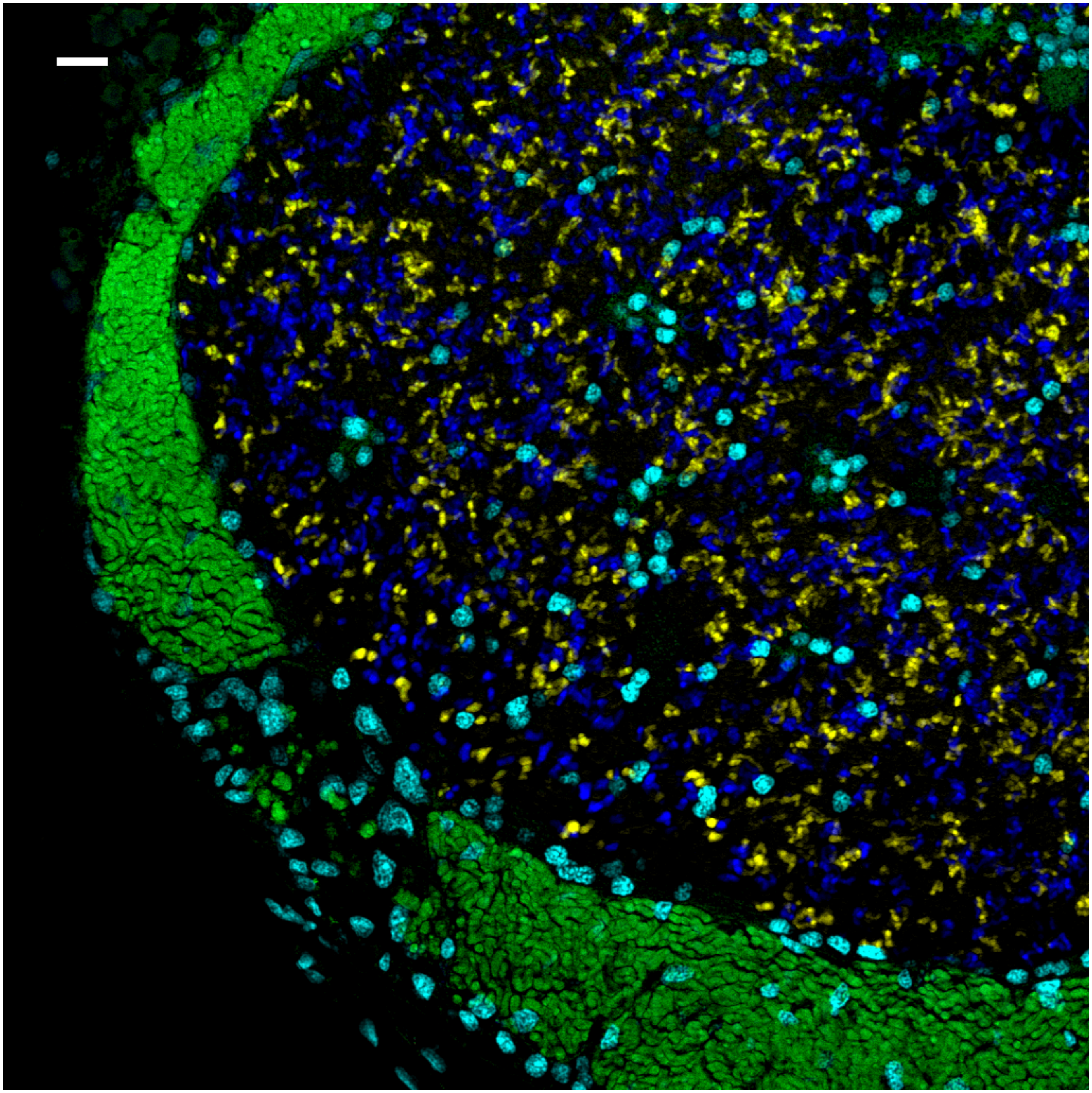


Figure S14. Fluorescent microphotograph of a cross-section of the bacteriome lobe of *Tettigades chilensis* specimen TETCHI. Cyan corresponds to Hoechst, universal DNA stain, and green represent the signal of fluorescently labeled probes that target 16S rRNA of *Sulcia*. Purple and yellow represent the signal of probes specific to rRNA of two distinct *Hodgkinia* variants: 16S rRNA of the lineage TETCHI2 and 23S rRNA of the lineage TETCHI4, respectively. Scale bar: 20 µm.

Supplementary Information – extended Material and Methods

1. Cicada material: specimen handling and an overview of molecular work

The adult cicada specimens used in this study were captured in Chile between 2006 and 2016. The detailed list is provided in Table S1. At each sampled site, captured specimens were sorted into morphospecies based on morphological characters. In many cases, multiple morphologically similar individuals were available from a single location, and they were preliminarily regarded as representatives of a single population of one species. Most specimens were directly preserved whole in 70% or 95% ethanol; others had their abdomens dissected and preserved in RNAlater, formaldehyde solution or glutaraldehyde solution, while the remainder of the body was placed in ethanol.

Representative specimens from all populations were compared against the morphological descriptions of the known species in the genus *Tettigades* (1, and references therein). The morphology-based identifications and the morphospecies boundaries were later confirmed using molecular methods.

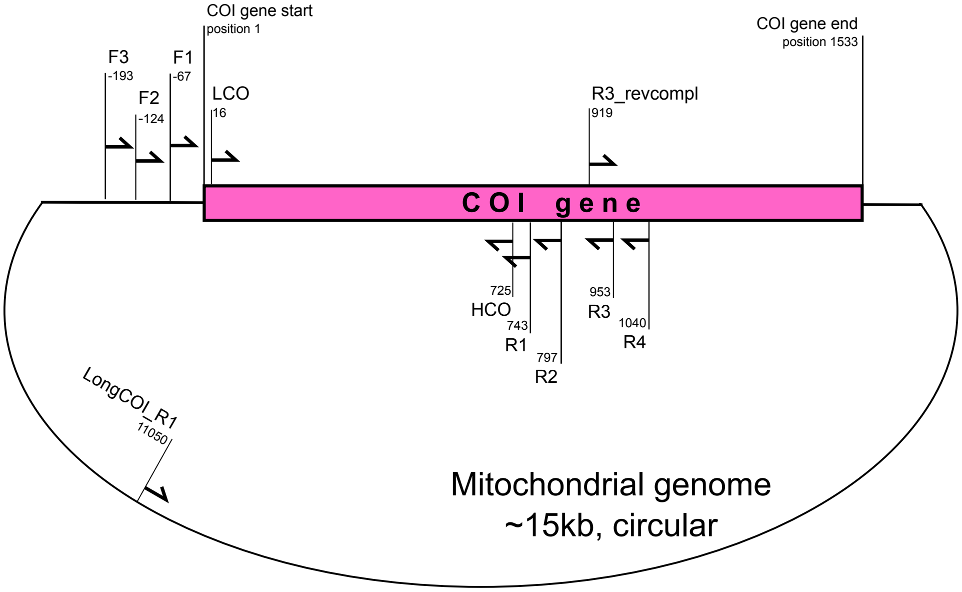
DNA was extracted either from dissected bacteriome tissue or from separated cicada legs. We used the DNeasy Blood and Tissue Kit (Qiagen Ltd.), and followed the manufacturer's standard protocol. The extracted DNA was used for amplification and sequencing of the mitochondrial cytochrome c oxidase I (*CO*I) gene of the hosts (see section 2). Using the combined morphological and COI sequence data, we selected nineteen divergent populations of *Tettigades*, and from each of them we designated one or two specimens for symbiont characterization. The phylogenetic placement of *Sulcia* was estimated by partially sequencing three protein-coding genes, RNA polymerase subunit B (*rpoB*), ribosomal protein L2 (*rplB*) and an outer membrane protein assembly factor BamA (*bamA*), selected as some of the most variable in the highly conserved genomes of *Sulcia* (see section 3). The diversity and phylogenetic relationships of *Hodgkinia* in experimental cicadas were estimated using amplicon sequencing of a 498bp region of *rpoB* gene, conserved in all *Hodgkinia* lineages from *Tettigades* spp. for which we had genomes sequenced (see sections 4 and 5). Phylogenetic analyses of these data (see section 8) helped us identify samples for metagenome characterization. Metagenomic libraries were prepared from bacteriome DNA of selected specimens using Illumina or NEB kits following the manufacturers' protocols, and sequenced on various Illumina platforms (Table S2). Metagenomic data analysis and genome annotation details are provided in section 6 and section 7, respectively. Bacteriomes of several of the field-collected *Tettigades* samples were used for ultrastructural observations using Transmission Electron Microcopy (TEM), and details are provided in section 9. Finally, bacteriomes of some specimens were used for fluorescence microscopy (see section 10).

2. Amplification and sequencing of the mitochondrial COI gene of cicadas

DNA barcoding – obtaining partial sequences of the mitochondrial cytochrome oxidase subunit I (COI) gene using universal primers, and comparing them across specimens in a collection, and against references in databases – has proven to be an effective way of surveying diversity in a sample of organisms (2). However, COI sequencing does work better for some taxa than for others. We were generally unable to obtain clean COI sequence traces when using universal barcoding primers developed by Folmer and colleagues (3), LCO1490 and HCO2198, and we developed a set of new primers based on previously sequenced mitochondrial genomes of four *Tettigades* species. These primers (TETbar_F1-F3, TETbar_R1-R4 – sequences below), used in various combinations, have resulted in clean sequence traces for approximately half of the specimens, but not for others. We concluded that these problems were likely due to high numbers of mitochondrial DNA copies in the nuclear genome of the cicada (numts) (4). To overcome these problems, we developed a two-step amplification protocol. In the first step, we used ultra high-fidelity polymerase Q5 (NEB) to amplify a ~11kb or a ~15kb region of the mitochondrial genome (assumed to be circular-mapping, as in the majority of other insects - see 5). This long PCR product, cleaned by digestion with Exonuclease I (NEB), was used as a template for the second step, where we attempted to amplify a much smaller portion of the gene using TETbar primers. The product of the second step was cleaned by digestion with Exonuclease I and Antarctic Phosphatase (NEB) and then submitted to Eurofins Genomics for Sanger sequencing using primer TETbar_R1. The sequence traces were aligned, checked for the presence of ambiguous peaks and trimmed so that the resulting sequence corresponded to positions 1-690 of the COI gene in *Tettigades ulnaria* specimen TETULN. In all cases, we verified whether the sequences had intact reading frames and no major changes in the amino acid sequence, which could indicate that a pseudogene was amplified. Using these approaches, we obtained clean and reliable sequences for all specimens but one, PL696.1. For that last sample, the product of the second-round PCR was cloned using JM109 competent cells (Promega), following the manufacturer's protocols. 20 randomly selected clones were sequenced. After discarding noisy sequences and those with frameshifts, we selected as representative for PL696.1 the cloned sequence that was a consensus of all sequences in the most abundant clade, that was also most similar to the sequence of *Tettigades lacertosa* PL675.

Details of primers and protocols are provided below.

***Tettigades*-specific COI amplification and sequencing primers:**
TETbar_F1 - ACATGTCAAAAGAACATTGTTCATTC
TETbar_F2 - GAATTTATTTCAAAATTGCAGTTTG
TETbar_F3 - GGCTTTAAGTTAATTAAACTATTATCC
TETbar_R1 - CCAGGTAAAATAAGAATATATACYTCAGG
TETbar_R2 - AATGATTCATTCCTACCTCTTTCTTG
TETbar_R3 - ACTTTAATACCTGTTGGTACAGC
TETbar_R4 - GTAAACAAAAACACGAATCCTAATG
LongCOI_R1 - GAGCCAGGTTGGTTTCTATC
TETbarR3_revcompl - CTATAATYATTGCTGTACCAACAGGT



Relative positions of primers used for amplification and sequencing of the barcoding region of the mitochondrial COI gene in *Tettigades* spp. Numbers represent positions of 5' ends of primers relative to the first annotated base of the COI gene in the mitogenome of *Tettigades ulnaria* (specimen TETULN).

Two-step PCR for amplification of COI: step 1 (long-range PCR)

Primer combinations:
TETbar_R1 & TETbarR3_revcompl
TETbar_F3 & LongCOI_R1

Master mix

Q5 polymerase (NEB)	0.2 µl
Q5 buffer	4.0 µl
40 mM dNTPs	0.4 µl
primer F @ 20 µM	0.5 µl
primer R @ 20 µM	0.5 µl
template	1.0 µl
water	to 20 µl

Cycling conditions:
30s @ 98°C; 8 cycles of 15s @ 98°C, 30s @ Annealing_Temp, 7-9min @ 72°C;
[The initial Annealing_Temp was 59-61°C, and it decreased by 0.5°C with each cycle]
25 cycles of 15s @ 98°C, 30s @ 55°C or 57°C, 7-9min @ 72°C;
2 min @ 72°C

Two-step PCR for amplification of COI: step 2 (specific PCR)

Primer combinations:
TETbar_F3 & TETbar_R1

Master mix

MyTaq HS Red Mix (Bioline)	12.5 µl
primer F @ 20 µM	0.5 µl
primer R @ 20 µM	0.5 µl
template	1.0 µl
water	to 25 µl

[As template, we used PCR product from the first step, after digestion with Exonuclease I (NEB)]

Cycling conditions
2min @ 95°C; 20 cycles of 15s @ 95°C, 30s @ Annealing_temp, 45s @ 72°C;
[The initial Annealing_temp was 60°C, and it decreased by 0.5C with each cycle]
25 cycles of 15s @ 95°C, 30s @ 50°C, 45s @ 72°C;
2min @ 72°C

In some cases, the annealing temperatures and the number of cycles were increased, resulting in a more ‘aggressive’ touchdown PCR conditions for those specimens where the product was not clean. We reasoned that starting the touchdown at a higher temperature should favor the specific COI product over potential pseudogenes.

3. Amplification and sequencing of protein-coding genes of *Sulcia*

The phylogenetic placement of the *Sulcia* symbiont of the experimental cicada specimens was estimated by partially sequencing three protein-coding genes, RNA polymerase subunit B (*rpoB*), ribosomal protein L2 (*rplB*) and outer membrane protein assembly factor BamA (*bamA*). These genes were selected as having some of the most variable stretches of sequence in the highly conserved genomes of *Sulcia*, and primers were designed using alignments of multiple previously sequenced *Sulcia* genomes. All three genes were amplified using touchdown PCR protocols with MyTaq HotStart Red Mix (Bioline Ltd.). Products were cleaned with 1.8x SPRI beads and sequenced in both directions. Sequences were trimmed to the aligned length of 1122bp (*rpoB*), 483bp (*rplB*) or 1134 bp (*bamA*).

***rpoB* primers:**
rpoB_F - TTA GTG GAT TCT GCT CCA AC
rpoB_R - TC TTC CTA CTT CTC CTA AAG AAT AGT
***rplB* primers:**
rplB_F - CAG GAG GTA GAA ATA ATT GTG GA
rplB_R - GGT CAA CTG GAT TCA TAG CT
***bamA* primers:**
bamA_F3 - AAG ATG AAA TCA TAT TCA GAG AAT TAA CA
bamA_R2 - TCA AGA GTT TTA TCC AAT CTA TAT GTT AGA

Master mix - all *Sulcia* genes:

MyTaq HS Red Mix (Bioline)	12.5ul
primer F @ 20 µM	0.625 µl
primer R @ 20 µM	0.625 µl
template	1.0 µl
water	to 25 µl

Cycling conditions - *rplB* and *rpoB*:
2min @ 95°C;
20 cycles of 15s @ 95°C, 30s @ Annealing_Temp, 45s @ 72°C;
[The initial Annealing_Temp was 58°C, and it decreases by 0.5°C with each cycle]
25 cycles of 15s @ 95°C, 30s @ 50°C, 45s @ 72°C;
2min @ 72°C

Cycling conditions - *bamA*:
As for rplB/rpoB, except that the Annealing_Temp at each cycle was 2°C higher

4. Sequencing of *Hodgkinia rpoB* amplicons

The diversity and phylogenetic relationships of *Hodgkinia* strains in experimental cicadas were estimated using amplicon sequencing of RNA polymerase subunit B (*rpoB*) gene. Our genomic analyses of 23 *Hodgkinia* lineages from six *Tettigades* species (see below) indicated that this protein-coding gene had been retained in genomes of all lineages and was among the most conserved. Primers were developed in highly conserved regions of the gene, which flanked a more variable 498bp region. These primers, complete with Illumina adapters, were used for the first round of PCR with ultra high-fidelity polymerase Q5 (NEB). The products, digested with Exonuclease I and Antarctic Phosphatase (NEB), were used for the second, indexing PCR, as described by Kircher and colleagues (6). The resulting amplicon libraries were pooled after rough quantification (comparison of band brightness on the agarose gel stained with ethidium bromide in presence of standards) and sequenced in a multiplexed 2 x 300bp Illumina Miseq lane.
Amplicon Library Preparation Protocol

Primers with Illumina adapters - first round of PCR
rpoB_3199R_P5a - ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCTRAGYTTAAYAAACGGATG
rpoB_3199R_P5b - ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGCTRAGYTTAAYAAACGGATG
rpoB_3199R_P5c - ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGCTRAGYTTAAYAAACGGATG
rpoB_3199R_P5d - ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCGCTRAGYTTAAYAAACGGATG

rpoB_2700Fb_P7 - GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATCGDTATTGCGMRGAGCTT

Template-specific primer sequences are underlined. Primers rpoB_3199R_P5a, _P5b, _P5c and _P5d differ from each other by the presence of a variable length insert (RED); the goal was to increase the nucleotide diversity across bases of the first read, aiding with cluster identification and improving sequence quality. These four primer variants were mixed at equimolar concentrations.

Indexing primers - second round of PCR

indexing_P5 - AATGATACGGCGACCACCGAGATCTACACnnnnnnnACACTCTTTCCCTACACGACGCTCTT
indexing_P7 - CAAGCAGAAGACGGCATACGAGATnnnnnnnGTGACTGGAGTTCAGACGTGT
A series of ns in primer sequence indicates a barcode; each sample had a distinct P5 and P7 barcode combination. Barcode sequences follow the lists in Meyer and Kircher (7) and Kircher *et al.* (6).

Master mix - first round of PCR (done in triplicate for each sample):

Q5 polymerase (NEB)	0.3 µl
Q5 buffer	6.0 µl
40 mM dNTPs	0.6 µl
primer F @ 20 µM	0.75 µl
primer R @ 20 µM	0.75 µl
template	2.0 µl
water	to 20 µl

Cycling conditions - first round of PCR:

30s @ 98°C;
27 cycles of 10s @ 98°C, 10s @ 60°C, 20s @ 72°C;
2min @ 72C

Triplicate samples were combined after the first round of PCR and prior to the ExoAP digestion.

Digestion of PCR products with Exonuclease I and Antarctic Phosphatase - Master mix:

ExoI enzyme (NEB)	0.2 µl
AP enzyme (NEB)	0.2 µl
AP buffer	1.0 µl
water	2.6 µl
PCR product	6.0 µl (2 µl from each of three replicate PCR products)

Incubate for 30 min @ 37°C + 15min @ 58°C

Use products diluted 3 fold as a template for a subsequent PCR.

Master mix - second round of PCR (indexing):

Q5 polymerase (NEB)	0.2 µl
Q5 buffer	4.0 µl
dNTPs	0.4 µl
primer F @ 5 µM	2.0 µl
primer R @ 5 µM	2.0 µl diluted
template	2.0 µl (3-fold dilution of ExoAP digest)
water	to 20 µl

Cycling conditions - second round of PCR (indexing):

30s @ 98°C;
6 cycles of 10s @ 98°C, 10s @ 65°C, 30s @ 72°C;

2 min @ 72°C

After the indexing step, 4 µl of each library was run on a 2% agarose gel stained with ethidium bromide, in presence of standards. The band brightness was estimated visually, and brightness scores were used to pool libraries to approximately equal DNA concentrations. Library pools were bead-cleaned, quantified using Qubit 2.0 and qPCR, and sequenced on a 2x300bp Illumina MiSeq lane.

5. Analysis of *Hodgkinia rpoB* amplicon data

The data were analyzed using mothur v. 1.37.4 (8). Reads were merged into contigs, which were then quality filtered. Because of very high read quality in the sequencing lane that included the majority of studied samples, we were able to quality-filter contigs very strictly: so that any with an average quality score of less than 25 in any 20-base window were discarded. Replicate specimens of *Tettigades undata* (Fig. 7B) were sequenced in a different MiSeq lane where read quality was lower, and we had to lower the filtering criteria. After identifying unique genotypes in the resulting filtered dataset, we discarded those represented by a single sequence only. We then aligned contigs against a set of *rpoB* sequences from the sequenced *Hodgkinia* genomes, again removing those that failed to align properly. After strict chimaera screening using UChime, we used the remaining reads for OTU clustering at the 97% identity level using the average neighbor algorithm. The specific commands for these steps are provided below.

The output files, including the alignment of all unique sequences, OTUs that each was assigned to, and the number of times each of them appeared in each of the libraries, were manually processed using custom Python scripts, Microsoft Office, and CodonCode Aligner v. 5.1.5. Briefly, trees were constructed for alignments of all unique genotypes from a given library. Then, in each clade / OTU, we identified those unique genotypes that were represented by much higher number of reads (more than 10-fold difference) than other unique genotypes in the same clade, and which were accompanied by multiple low-abundance genotypes that differed at a single nucleotide positions. These abundant unique genotypes, typically one but sometimes two per OTU, were identified as error-free *rpoB* sequences of *Hodgkinia* lineages present in a given cicada specimen. We verified this approach by independently processing replicate specimens from twelve populations and comparing the identified sequences. We also verified the results by comparing sequences obtained using this method with gene sequences from sequenced genomes of all *Hodgkinia* lineages from five cicada specimens.

Mothur commands used

```
######## Set working directories
set.dir(input=..../20160608_MiSeq_run, output=..../20160608_MiSeq_run)

######## Assemble forward and reverse reads
make.contigs(file=samples.txt, processors=32)

######## Replace all hyphens in read and sample names with underscores
system(sed -i 's/-/_/g' samples.contigs.groups)
system(sed -i 's/-/_/g' samples.trim.contigs.fasta)
system(sed -i 's/-/_/g' samples.trim.contigs.qual)

######## Extract libraries for analysis from a larger sample set
get.groups(group=samples.contigs.groups, fasta=samples.trim.contigs.fasta, groups=TETAUR-
TETCHI-TETULN-TETUND-.....)
list.seqs(group=samples.contigs.pick.groups)
get.seqs(accnos=samples.contigs.pick.accnos, qfile=samples.trim.contigs.qual)

######## Quality-trimming sequences, removing all those with poor-quality fragments
trim.seqs(fasta=samples.trim.contigs.pick.fasta, oligos=primers_to_trim.oligos,
qfile=samples.trim.contigs.pick.qual, minlength=450, maxlength=550, maxambig=0,
maxhomop=10, qwindowsize=20, qwindowaverage=25, pdiffs=2, processors=32)
```

```
    ### Note: these very strict filtering criteria were only possible because of very
high read quality in the first lane.
    ### For the second lane that included replicate T. undata libraries, more relaxed
trimming criteria were used.
list.seqs(fasta=samples.trim.contigs.pick.trim.fasta)
get.seqs(accnos=samples.trim.contigs.pick.trim.accnos, group=samples.contigs.pick.groups)

######## Pick unique sequences
unique.seqs(fasta=current)

######## Generate count_table = a table with information on the number of times each
unique sequence appears in each library
count.seqs(name=current, group=current)

######## Discard singleton sequences
split.abund(fasta=current, count=current, cutoff=1)

######## Align sequences against rpoB reference (23 sequences, alignment = 505bp)
align.seqs(fasta=current, reference=rpoB_references.fasta, processors=4)

######## Remove sequences that did not align correctly
screen.seqs(fasta=current, count=current, start=1, end=505, minlength=470)

######## Chimera filtering using UChime - strict protocol
chimera.uchime(fasta=current, reference=self, count=current, dereplicate=f, mindiv=0.35,
processors=16, minh=0.5, xn=3)

######## Remove chimeric sequences
remove.seqs(accnos=current, fasta=current, count=current)

######## Remove
filter.seqs(fasta=current, vertical=T, trump=.)

######## Computing pairwise distance matrix, OTU picking
dist.seqs(fasta=current, processors=16, cutoff=0.20)
cluster(column=current, count=current, cutoff=0.20, method=average)

######## Binning sequences - 99% OTUs
bin.seqs(list=current, fasta=current, label=0.01)
make.shared(list=current,count=current, label=0.01)
```

6. Metagenome sequencing and assembly

We sequenced bacteriome metagenomes of five species from the genus *Tettigades*, as well as an outgroup, *Chonosia crassipennis* specimen CHOCRA (Table S2) using various Illumina platforms. Metagenomic reads were quality-trimmed using Trim Galore! (https://github.com/FelixKrueger/TrimGalore) and merged using pear v. 0.9.6 (9). Subsequently, merged and unmerged reads were used for assembly with SPAdes versions 3.1.1 and 3.7.0 (10), which had been compiled so that kmers beyond the standard limit of 127 could be used. Initially, all reads from a given library were used for assemblies with the maximum kmer size of 191bp. Subsequently, we identified scaffolds with significant similarity to the previously sequenced genomes of *Hodgkinia*, *Sulcia* and mitochondrion of *Tettigades ulnaria* using blastn and promer v. 3.0 (11). These contigs were used as references for read mapping using bwa v. 0.7.12-r1039 (12), with settings modified so that only reads with very high similarity to references would map. These mapped reads were extracted using SamToFastq (Picard Tools - https://broadinstitute.github.io/picard/), merged again using pear, and then used for SPAdes assemblies with maximum kmer size of 245. Gaps between scaffolds were closed using PCR and Sanger sequencing. Also, we have amplified and sequenced with a set of universal primers the complete rRNA operons of all *Hodgkinia* lineages from all newly characterized *Tettigades* spp., and complete *rpoB*-*rpoC* operons and several other conserved genes from two pairs of recently diverged genomes, TETCHI1a-TETCHI1b and TETLON2a-TETLON2b. Finally, we used a set of PCR reactions with long-range, high-fidelity polymerase Q5 (NEB) to verify alternative arrangements of some genomes. In all cases, the quality of the final genomic sequences was verified by mapping reads and the manual inspection of the alignments using Tablet (13).

7. Symbiont genome annotation and comparison

The genomes were analyzed and illustrated using a set of custom Python and Processing scripts. Annotation was conducted by recursive searches for a manually curated set of alignments of protein-coding, rRNA and ncRNA genes from all previously characterized *Hodgkinia* or *Sulcia* lineages using HMMER 3.1b2 (14). Based on the length of the longest ORF relative to the reference ORF in TETULN genome, genes were classified as functional (>85%), putative pseudogenes (>60%), or pseudogenes. Any open reading frames of at least 300 nucleotides that had not been annotated by the script were manually searched using hmmer and blastx/tblastx against UniProt and NCBI databases. All genes previously annotated as "hypothetical" or unannotated (15, 16) were carefully manually compared against the top hits in other microorganisms using blastp (https://blast.ncbi.nlm.nih.gov) and HMMER (https://www.ebi.ac.uk/Tools/hmmer), resulting in the discovery of additional genes.

Reference-based annotations of rRNA genes were supplemented by rRNA searches using RNAmmer v. 1.2 (17), and tRNA searches using tRNAscan-SE v. 1.4 (18). Alignments of all genes classified as functional were done using mafft v. 7.221 (19). In case of protein-coding genes, alignments were conducted in protein space and reverse-translated to nucleotide space.

8. Phylogenetic analyses

Phylogenies of the cicadas with bacteriome metagenomes sequenced, as well as their symbionts, were based on unambiguous alignments of all genes other than tRNA that had been classified as functional in all genomes characterized. These sets included, respectively, 15 genes from mitochondrial genomes (total alignment length 13,194 bp), 12 genes shared across the genomes of all *Hodgkinia* lineages (total alignment length 24,401 bp), and 230 genes found across *Sulcia* genomes (total alignment length 238,488 bp). The alignments were divided into partitions corresponding to three codon positions and to RNA genes as the fourth partition, after verifying using PartitionFinder2 (20) that this partitioning scheme provided a good fit to the data. Phylogenetic analyses were conducted using RAxML v. 8.2.9 (21) assuming GTR+GAMMA model, and with 100 rapid bootstrap replicates.

The resulting multi-gene mitochondrial, *Hodgkinia* and *Sulcia* trees were used to constrain phylogenies for a larger set of samples, which were based, respectively, on partial sequences of COI (690 bp), *rpoB* (498 bp) or on a concatenation of partial sequences of three *Sulcia* genes (*bamA* – 1134 bp; *rplB* – 483bp; *rpoB* – 1122bp; total length 2739 bp). In all cases, we used RAxML v. 8.2.9 with GTR+GAMMA model, partitions corresponding to three codon positions, and using 1000 bootstrap replicates.

9. Light and transmission electron microscopy (TEM) of bacteriome tissue

Partially dissected abdomens of males and females of the examined cicadas species were fixed in 2.5% glutaraldehyde in 0.1 M phosphate buffer (pH 7.4) at 4°C for three months. After four washes with phosphate buffer with addition of 5.8% sucrose, the bacteriomes were fully dissected and postfixed in 1% osmium tetroxide in 0.1 M phosphate buffer (pH 7.4). Then, the samples were rinsed in cold water and dehydrated in ethanol series (30%-100%) and then acetone, before embedding in epoxy resin Epon 812 (Serva, Germany). Semi-thin sections (1 µm thick) were stained with 1% methylene blue in 1% borax and analyzed and photographed under light microscope Nikon Eclipse 80i. Ultrathin sections (90 nm thick) for TEM studies were contrasted with saturated solution of lead citrate and uranyl acetate and examined using the Jeol JEM 2100 (Jeol, Japan) electron microscope.

10) Fluorescent microscopy of bacteriome tissue

Dissected bacteriome of single individuals of *T. chilensis* individual TETCHI, the same which was used for bacteriome metagenome sequencing, was also used for fluorescent microscopy. Originally preserved in ~90% ethanol, the tissue was rehydrated, fixed in 4% formaldehyde, dehydrated through 1 hr incubations in 80%, 90% and 100% ethanol, then cleared in methylscylate for 2 x 1 hr and embedded in paraffin under vacuum, for 2 x 1 hr. Paraffin blocks were sectioned to 5-10 µM. Thin sections were de-paraffinized in xylene and 100% ethanol (three washes in each) and then hydrated in tap water. Subsequenctly, we applied hybridization buffer containing 12.5% dextran sulfate, 2.5X SCC, 10ng/uL ssDNA, 0.25% BSA, as well as 1.5ug/uL Hoechst 33258, fluorescently labeled probes at 200 nM, and unlabeled helper oligos at 2 µM. The probes are listed below; Hodg302 and Sulc664 were modified from previous studies (16, 22), others were developed during the current study. Fluorescently labelled probes were used at a concentration of 200 nM, helper oligos were used at a concentration of 2 µM each. Hybridization was conducted overnight at 37°C, except that 1h into hybridization we applied heat shock (1 min @ 90°C); this was thought to denaturate rRNA and make it more accessible to probes which should have penetrated tissue by that time. After hybridization, slides were then washed with 2X SCC three times at 37°C over the course of 1 hr, and preserved with FluorSave (CalbioChem). Imaging was done on an Olympus FV 1000 IX inverted laser scanning confocal microscope with 20X air lens and 63X oil-immersion lens.

Probe name	**Sequence**	**Fluorophore**
TETCHI2_Cy3	CTGCTGTCGCTATTCG	Cy3
TETCHI2_Rhelper	ACGACTTCACCCCAGTTATCAAC	unlabelled (helper oligo)
TETCHI2_Lhelper	GTTTGCGATAGCTTAAAACAAAGCT	unlabelled (helper oligo)
TETCHI4_Cy5	GCAATGACATCGCAAAA	Cy5
TETCHI4_Rhelper	AACCTTTAGGCTATTTCCCGTT	unlabelled (helper oligo)
Hodg302_Cy5	CCAATGTGGCTGRCCGT	Cy5
Hodg302_Lhelper	CTCCCAGACCAGCTATAGATCRTCGCC	unlabelled (helper oligo)
Hodg302_Rhelper	CCGTAGAAGTTTGGGCCGTGTCTCAGT	unlabelled (helper oligo)
Sulc664_TF2	CCACACATTCCAGTTACTCC	Tide Fluor 2
Sulc664R_Lhelper	CCTCACTCTAGTTTATCAGTATCAATAGCACTT	unlabelled (helper oligo)
Sulc664R_Rhelper	GTTCTGTGTGATCTCTATGCATTTCACCGCT	unlabelled (helper oligo)

References

1. Torres BA (1958) Revision del genero "Tettigades" Amy. et Serv. (Homoptera-Cicadidae). *Revista del Museo de La Plata, Nueva Serie* 7:51–106.
2. Hebert PDN, Cywinska A, Ball SL, & deWaard JR (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B. Biol. Sci.* 270(1512):313-321.
3. Folmer O, Black M, Hoeh W, Lutz R, & Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 35(5):294-299.
4. Hazkani-Covo E, Zeller RM, & Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* 6(2):e1000834.
5. Cameron SL (2014) Insect mitochondrial genomics: implications for evolution and phylogeny. *Annu. Rev. Entomol.* 59:95-117.
6. Kircher M, Sawyer S, & Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3.
7. Meyer M & Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* 2010(6):pdb.prot5448.

8. Schloss PD*, et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75(23):7537-7541.
9. Zhang J, Kobert K, Flouri T, & Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30(5):614-620.
10. Bankevich A*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19(5):455-477.
11. Kurtz S*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol.* 5(2):R12.
12. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754-1760.
13. Milne I*, et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* 14(2):193-202.
14. Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comp. Biol.* 7(10):e1002195.
15. McCutcheon JP, McDonald BR, & Moran NA (2009) Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. *PLoS Genet.* 5(7):e1000565.
16. Van Leuven JT, Meister RC, Simon C, & McCutcheon JP (2014) Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. *Cell* 158(6):1270-1280.
17. Lagesen K*, et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35(9):3100-3108.
18. Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25(5):955-964.
19. Katoh K & Standley DM (2013) MAFFT Multiple Sequence Alignment Software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30(4):772-780.
20. Lanfear R, Calcott B, Ho SYW, & Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29(6):1695-1701.
21. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312-1313.
22. Bennett GM & Moran NA (2013) Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect. *Genome Biol. Evol.* 5(9):1675-1688.