

Rickettsial DNA invasions and a scrambled rRNA cluster with a trans-splicing group I intron: The highly unorthodox mitogenome of the fern *Haplopteris ensiformis*

Simon Zumkeller (✉ zumkellersimon@gmail.com)

University of Bonn <https://orcid.org/0000-0002-1630-2439>

Monika Polsakiewicz

Universität Bonn

Volker Knoop

University of Bonn <https://orcid.org/0000-0002-8485-9423>

Article

Keywords: Monilophyte mitogenome, Pteridaceae, Group I introns, Group II introns, Trans-splicing, C-to U and U-to-C RNA editing, Lateral and Horizontal Gene Transfer, Rickettsiales

Posted Date: October 12th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-2152556/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 Rickettsial DNA invasions and a scrambled rRNA cluster with
2 a *trans*-splicing group I intron: The highly unorthodox mitogenome
3 of the fern *Haplopteris ensiformis*

4 Simon Zumkeller, Monika Polsakiewicz and Volker Knoop*

5 IZMB – Institut für Zelluläre und Molekulare Botanik, Abteilung Molekulare Evolution,
6 Universität Bonn, Kirschallee 1, D-53115 Bonn

7 ORCID IDs:

8 Simon Zumkeller: 0000-0002-1630-2439

9 Volker Knoop: 0000-0002-8485-9423

10 *Corresponding author

11 Email: volker.knoop@uni-bonn.de

12 Tel +49 228 73 6466, Fax +49 228 73 6467

13 **Key words:** Monilophyte mitogenome, Pteridaceae, Group I introns, Group II introns, *Trans*-splicing,
14 C-to U and U-to-C RNA editing, Lateral and Horizontal Gene Transfer, Rickettsiales

Summary

Plant mitochondrial genomes can be complex owing to highly recombinant structures, lack of gene synteny, heavy RNA editing and invasion of chloroplast, nuclear or even foreign DNA by horizontal gene transfer (HGT). Leptosporangiate ferns remained the last major plant clade without an assembled mitogenome, likely owing to a demanding combination of the above. We here present both organelle genomes of *Haplopteris ensiformis* as a first leptosporangiate fern. More than 1,400 events of C-to-U RNA editing and over 500 events of reverse U-to-C edits affect its organelle transcriptomes. The *Haplopteris* mitogenome has a rich gene complement lacking only the *ccm* gene suite but is highly unorthodox, indicating extraordinary recombinogenic activity. Although eleven group II introns known in disrupted *trans*-splicing states in seed plants exist in conventional *cis*-arrangements, a particularly complex structure is found for the mitochondrial *rrnL* gene, which is split into two parts needing reassembly on RNA level by a novel *trans*-splicing group I intron. Finally, and aside from ca. 80 chloroplast DNA inserts that complicated the mitogenome assembly, the *Haplopteris* mtDNA features as an unparalleled idiosyncrasy 30 variably degenerated protein coding regions from Rickettiales bacteria indicative of heavy bacterial HGT on top of tRNA genes of chlamydial origin.

Introduction

The conserved structure of chloroplast genomes (plastomes) and the stoichiometric dominance of chloroplast (cpDNA) over nuclear and mitochondrial (mtDNA) in total plant nucleic acid preparations have led to a tremendous increase of available complete plastome sequences with the advent of Next Generation Sequencing (NGS) technologies. The number of completed plant mitochondrial genome sequences (mitogenomes) is much lower in contrast, most notably in vascular plants (tracheophytes). This is largely due to the much more variable and complex mitogenome structures in tracheophytes¹⁻⁵, which hitherto left complete chondrome assemblies altogether missing for the large clade of leptosporangiate ferns among the monilophytes.

Complex, recombining mitogenomes clustered with repeat sequences and affected by the lateral invasion of chloroplast DNA, or even by horizontal gene transfer (HGT) from other species, have arisen independently in the four large tracheophyte clades: the angiosperms, the gymnosperms, the monilophytes and the lycophytes. The lycophytes as the evolutionary oldest of the four clades of extant vascular plants reflect this most clearly: The mitogenome in the club moss *Phlegmariurus squarrosus* is a circular DNA with a rich gene complement and even retaining several ancestral gene syntenies of the circular mtDNAs in bryophytes⁶. The mtDNAs of the quillwort *Isoetes engelmannii* or the spike moss *Selaginella moellendorffii*, in contrast, are strongly depauperated in gene content and heavily affected by recombination leading to complex and coexisting arrangements of coding islands embedded between repeated sequences^{7,8}. Similarly, moderately compact and gene-rich circular mitogenomes are present in the gymnosperms *Cycas taitungensis*⁹ and *Ginkgo biloba*¹⁰. In contrast, the mitogenomes of *Taxus cuspidata*¹¹ and *Welwitschia mirabilis*¹⁰ are reduced in gene content and this even despite size increase to nearly 1,000 Kbp in the latter case. As reflected from a collection of mitochondrial scaffolds and chromosomal assemblies, respectively, yet much larger mitogenomes are present in the conifer I clade in *Picea* species^{12,13} or *Larix sibirica*¹⁴.

Among angiosperms, the mtDNA of the magnoliid *Liriodendron tulipifera* represents an ancestral, gene rich state¹⁵. Other flowering plant species, however, have complex mtDNAs of enormous sizes

exceeding 1,000 KBp, are fragmented into multiple co-existing mitochondrial chromosomes, affected by massive lateral gene transfer (LGT) from the cpDNA or by horizontal gene transfer (HGT) from other species or display combinations of those features to variable degrees. The large mitogenomes exceeding 1 MBp in the cucumber family^{16,17}, the diverse multipartite mtDNAs in the genus *Silene*¹⁸ or the mtDNA of the isolated flowering plant *Amborella trichopoda* that is heavily affected by HGT from diverse species¹⁹ are prime examples.

Among monilophytes (ferns *sensu lato* including the horsetails), the second largest group of tracheophytes behind the angiosperms, the situation is less clear since only the two complete chondrome sequences of *Ophioglossum californicum* (adder's tongue) and the whisk fern *Psilotum nudum* have been determined²⁰. Although the two taxa represent sister clades among the eusporangiate ferns, a paraphyletic grade at the base of extant fern taxa, their mitogenomes reflect differences indicating a dynamic evolution of mtDNAs also among the monilophytes. Aside from slight differences in gene and intron complement, the *O. californicum* mtDNA maps as a single circular mtDNA whereas two separate circular chromosomes exist in *P. nudum*²⁰.

Most of extant fern diversity with more than 10,000 species resides in the leptosporangiate ferns, however. Studies on selected mitochondrial loci among ferns revealed interesting dynamics of group II introns in their mtDNAs^{21,22}. The leptosporangiate fern family of Pteridaceae (Polypodiales) proved to be particularly interesting in a study of mitochondrial group II intron gain, loss and coevolution scenarios²², also with respect to the concomitant evolution of C-to-U and "reverse" U-to-C RNA editing, which is abundantly present in the endosymbiotic organelles of leptosporangiate ferns^{23,24}. The Pteridaceae represent a large fern family comprising some 1,150 species in 45 genera, placed into at least five sub-groups and potential sub-families: the cryptogrammoid, the adiantoid-vittarioid, the cheilanthoid-hemionitidoid, the ceratopteroid-parkerioid, and the pteroid ferns. Habitats occupied by these sub-groups are equally diverse, as they range from terrestrial, including epipetric and epiphytic, to even aquatic lifestyles.

We chose *Haplopteris ensiformis* among the epiphytic, vittarioid “shoestring ferns” for a detailed analysis of its two organelle genomes and transcriptomes. We found a typical, conserved chloroplast genome structure in *H. ensiformis* but identified a highly unorthodox mitogenome characterized by numerous active and inactive repeat sequences and a massive insertion of chloroplast DNA. Transcript maturation of the comparatively rich mitochondrial gene complement, lacking only *ccm* genes for cytochrome maturation, involves splicing of 24 group II and four group I introns and abundant C-to-U and U-to-C RNA editing at nearly 2,000 sites. The most surprising novelties of molecular evolution in a plant mitogenome include a unique novel *trans*-splicing group I intron in the large ribosomal rRNA and, most notably, extended stretches of bacterial, *Rickettsia*-like DNA in the *Haplopteris* mtDNA.

Results

Our choice of the “shoestring fern” *Haplopteris ensiformis* as a first leptosporangiate fern for complete assembly of both organelle genomes was based on pronounced variability in mitochondrial RNA editing and intron (co-)evolution in the monilophyte family Pteridaceae and the sub-family Vittarioideae in particular²². We used Next Generation Sequencing (NGS) data to assemble the chloroplast (cpDNA) and mitochondrial genomes (mtDNA) of *H. ensiformis*, accompanied by RNA-seq transcriptome analyses to study RNA processing with a special focus on intron splicing and RNA editing. As expected, chloroplast DNA reads dominated in the NGS data with read coverages of ca 1,800 to 4,200 and allowed the straightforward assembly of the *Haplopteris ensiformis* cpDNA. The mitochondrial DNA, in contrast, revealed overall lower and much more variable read coverages and its assembly was highly complicated by a multitude of repeats, long intergenic stretches and the insertions of foreign DNA and required multiple independent PCR amplifications for verification and complete assembly.

The Haplopteris ensiformis plastome and its well-resolved RNA editome

The *Haplopteris ensiformis* chloroplast DNA (148,805 bp) reveals a typical conserved circular plastome structure with a large (80,986 bp LSC) and a small (20,773 bp SSC) single copy region separated by a pair of inverted repeats (IRs of 23,523 bp each). The chloroplast genome carries 116 genes widely

conserved in other land plants, 85 of which encode proteins, including the recently characterized *ycf94* gene, four rRNAs and 27 tRNAs ([Fig. 1A](#)). Likewise, the *H. ensiformis* cpDNA contains a set of 20 conserved introns. Our accompanying transcriptome analysis confirmed functional splicing for all of them.

One striking structural difference concerns the presence of two morffo elements, “mobile ORFs in fern organelles”²⁵ in the chloroplast genome of the related species *Haplopteris elongata* ([Fig. 1B](#)). Our data suggest their secondary loss in the now determined cpDNA of *H. ensiformis* rather than an independent gain in *H. elongata* as we will discuss below in the context of the numerous cpDNA inserts that we identified in the now determined *H. ensiformis* mitogenome.

Modern RNA-Seq technologies allow for detection also of low-rate RNA editing sites with reasonable precision and any numbers of reported edits for an organelle transcriptome should nowadays ideally be accompanied by threshold criteria for their detection. Likewise, instead of the frequently used terms “complete” or “partial” editing, the percentage of detected base conversions by RNA editing should be given for the respective editing sites. The generally high coverage of RNA-Seq reads for the chloroplast transcriptome (mostly above 1,000 x and here reaching coverages of up to 250,000 x for the *psbE* gene in the case of the *H. ensiformis* cpDNA) allows for a detailed evaluation of RNA editing events with high precision, allowing for three digits behind the decimal point. Using stringent criteria for DNA and RNA read qualities (see Materials and Methods) we could identify 443 sites of chloroplast RNA editing covered by at least 100 RNA reads and RNA editing frequencies of at least 1.0 % ([Suppl. Tab. 1](#)).

We here use our previous nomenclature proposal for unequivocal labeling of RNA editing events²⁶, indicating the affected locus, the nucleotide resulting from C-to-U or reverse U-to-C editing (eU or eC), the position and, for the majority of edits within protein coding regions, the codon meaning before and after the edit ([Suppl. Tab. 1](#)). The careful analysis of RNA editing events contributes significantly to the identification of functional genes in the organelles or the dismissal of others as pseudogenes,

notably plant species like *H. ensiformis* featuring abundant C-to-U and reverse U-to-C RNA editing at the same time. A case in the point is the small reading frame *ycf94/orf51* of hitherto unknown function between *rps16* and *matK*, for which we here find even higher rates of RNA editing at two important sites (*ycf94eU2TM* and *ycf94eU50PL*) of 59% and 74%, respectively, than reported previously for other species^{27,28}. *Vice versa*, we consider *rps16* a degenerating pseudogene in *H. ensiformis* as we could not confirm the expected removals of stop codons by reverse U-to-C editing.

The range of observed editing efficiencies extends from 99.1 % for the codon sense-changing editing event *psbBeU116SL* down to 1.0 % for edits *accDeU657SS* (silent), *petBi6g2eU345* (intron) or edit *rps2eU349R**, which unexpectedly introduces an early stop codon in the *rps2* coding sequence. Such edits at low frequency are likely “collateral” effects owing to lacking specificity of the chloroplast RNA editing machinery. The same holds true for most codon sense-changing edits that are unexpected (as they do not restore conserved codon identities) and likewise show only inefficient editing with low frequency (Suppl. Tab. 1).

Most of the detected non-silent RNA editing sites in the chloroplast coding sequences, however, confirm expectations for restoring conserved amino acid positions very well and are efficiently edited and, *vice versa*, we find low frequencies of editing nearly exclusively in silent or non-coding positions like 5'- or 3'-UTRs (Suppl. Tab. 1). However, exceptions exist: Prime example for efficient RNA editing events in non-coding regions, which could have been missed altogether in typical RT-PCR-based studies focusing on coding sequences, are cytidine-to-uridine conversions *petBi6g2eU478* and *rps12i346g2eU80* in the respective introns within *petB* and *rps12* of 94.0 % and 90.2 %, respectively. Conversely, we observe strikingly low efficiencies in many cases of reverse U-to-C editing removing stop codons, for example *rpoC2eC232*Q* reconstituting a glutamine codon in the *rpoC2* reading frame is edited with only 17.0 % efficiency in the transcript population.

Ten start codons and one stop codon are created by C-to-U editing and altogether 26 stop codons are removed by reverse U-to-C RNA editing in the chloroplast gene transcripts (Fig. 1, Suppl. Tab. 1).

We here use the *accD* gene as a somewhat less conserved protein coding region as an example for discussion (Fig. 2). Codon-changing edits confirm predictions very well with editing frequencies between 61.0% for accDeC730SP and 94.1% for accDeU779SL. Unpredictable edits in the UTRs or silent position accDeU657SS are edited much less efficiently except for accDeU-1 right upstream of the start codon created by editing (Fig. 2B). An intriguing case is editing site accDeU625HY inefficiently edited to 6.4%, for which a histidine or tyrosine is found variably in other taxa. An additional reverse edit is predicted for position accDeC580FL but remained unconfirmed. Note that in most fern cpDNA database entries, just two RNA editing sites are arbitrarily postulated to create an intact *accD* reading frame with a start codon edit in position 2 and to remove a stop codon in position 772.

Given that the organelle transcriptomes of *Haplopteris ensiformis*, and especially the mitochondrial transcriptome (see below), proved to be new examples for abundant C-to-U and U-to-C editing, we here use the opportunity to introduce a nomenclature amendment addressing the complex issue of multiple editings in individual codons. We suggest to indicate the individual and cumulative effects on codon meaning after and before a pipe symbol (|), respectively, and additional silent edits by an underline symbol (Suppl. Fig. S1). As an example we consider edits atpBeU1381PL|PS and atpBeU1382PL|PL changing a proline (P) codon identity in the *atpB* transcript. The first position edit takes place with 97.9% efficiency to expectedly reconstitute a conserved serine (S) codon whereas the unexpected additional 2nd position edit causes a change into a leucine (L) codon, although with only 1.7% efficiency. Yet more complex is the example of a CCC proline codon in the *rpoB* gene edited with different frequencies in all three positions. Edit rpoBeU662PF|PL in the second codon position causes the expected change towards a conserved leucine codon with 66.3% efficiency. However, unexpected edit rpoBeU661PF|PS in the first codon position with only 1.2% efficiency would, considered alone, cause a change to serine. In combination with the edit in the second position the codon is changed to a phenylalanine (F) codon (Suppl. Tab. 1, Suppl. Fig. 1). Moreover, these two non-silent edits are accompanied by a 3rd position edit with 36% efficiency (rpoBeU663PP_FF), which is silent for any of the 4 possible codon identities (P, L, S or F).

Assembly of the highly complex *Haplopteris ensiformis* mitogenome

The assembly of the *Haplopteris ensiformis* mitogenome turned out to be very demanding owing to a combination of several factors, which we will address in separate paragraphs below. The mtDNA sequence reads were not only ca. 10-fold less abundant than those of the cpDNA (on average ca. 150 x) but also much more variable in coverage (ca. 70 x to 600 x). We relied on parallel transcriptome analysis to verify authentic native mitochondrial genes characterized by RNA editing. Evident mtDNA contig assemblies ran into numerous repeated sequences, generally represented with higher read coverage or into extended insertions of laterally transferred chloroplast DNA fragments, a typical feature of vascular plant mitogenomes. Finally, we found numerous surprising similarities with *Rickettsia*-type bacterial genomes indicating multiple horizontal gene transfer (HGT) beyond the previously identified chlamydial tRNA genes in early branching tracheophytes ²⁹.

Multiple recombination breakpoints allow for a huge spectrum of alternative, and likely co-existing, mitogenome arrangements with variable stoichiometries. For clarity, we chose to assemble twelve mtDNA contigs (A to L, ranging in sizes from 2,646 to 77,705 bp) comprising the full mitochondrial sequence complement into nine circular mtDNA chromosomes (chr1 to chr9) as separate GenBank database accessions (OM867545-OM867553). These nine circular chromosomes, however, likely represent only a substoichiometric minority of the truly existing mtDNA molecules owing to the numerous recombination breakpoints (Figs. 3 and 4). In total, 32 recombination breakpoints could be identified. These recombination breakpoints are labeled with 'r', the respective contig and consecutive numbers (rA1-rL2). We here display chromosome 5 comprising contigs A-C ([Fig. 3](#) and chromosomes 1-4 and 6-9, variably integrating contigs D-K into the former ([Fig. 4](#)). Among these, contig A features prominently as it is flanked at its ends by repeat R9170 in inverted orientations, the largest repeat sequence that we identified in the *H. ensiformis* mitogenome (Fig. 3). Repeat R9170 carries the upstream part of the ribosomal rRNA gene cluster (Fig. 3), hence distantly reminding of the typical IRs in chloroplast genomes (see Fig. 1).

Repeats and Recombination in the *Haplopteris ensiformis* mitogenome

Other than by the many chloroplast DNA inserts that we will discuss below, the assembly of the *Haplopteris ensiformis* mtDNA was complicated by numerous repeated sequences of different sizes. Ancestral mitochondrial gene synteny is widely eradicated in *H. ensiformis* owing to numerous recombination events – the *rps19-rps3-rpl16* gene continuity on contig F is one remaining exception (Fig. 4). While single-copy sequence contigs with mitochondrial genes had average NGS DNA read coverages around 150-fold (albeit with a broad distribution), identical repeat sequences mostly had coverages exceeding 300-fold. For clarity and discussion, we have labelled and annotated repeats with ‘R’ followed by the number of nucleotides, also in the corresponding GenBank accessions.

We carefully checked on recombinational activity across repeats with a “template-switch-avoiding” tsa-PCR strategy and examples are shown in figure 5. Repeat R596 (Fig. 5A) is a particularly intriguing example as it is not intergenic but shared as an identical sequence present in domains I of group II introns *nad5i1242g2* and *rrnLi833g2*, located on contigs H and F, respectively (Fig. 4). R596 had a coverage of ca. 500 x whereas the flanking single-copy contigs with *nad5* and *rrnL* coding sequences had coverages of ca. 150 x and 350 x, respectively (Fig. 5A), indicating a non-equilibrium and different stoichiometries of the *nad5* and *rrnL* gene copies. Naturally, active recombination would cause dysfunctional chimeric *nad5/rrnL* genes in this case. Exploring potentially active recombination across R596, we confirmed the gene continuities for *nad5* and *rrnL* but found only very weak evidence for active recombination potentially giving rise to the two reciprocal arrangements (Fig. 5A).

However, similar tsa-PCR approaches run across other (intergenic) repeats indicated (stronger) active recombination across the respective repeated sequence resulting in more than only two co-existing, or at least strongly dominating, conformations of the flanking single-copy sequences. Examples are shown for R203 and R513 (Fig. 5B). Whereas all combinations of flanking sequences appear to co-exist for R203, the arrangements A-C and B-D appear to strongly dominate for R513 while arrangement A-D is stoichiometrically under-represented and arrangement B-C even remains undetected (Fig. 5B).

As an additional complication, some recombination points at the end of repeats were in in very close proximity resulting in complex “combined” repeats. We here show results for R180, one copy of which is in group II intron nad2i709g2 of the *nad2* gene (Fig. 5C). In another location, R180 is flanked on one side by consecutive repeats R137 and R259 resulting in alternative pathways to three different endpoints (D, E and F in Fig. 5C). Endpoints D and E result in intact *atp4* and *atp6* gene copies, respectively, whereas the alternative arrangement A-F creates pseudogene fragments for both genes downstream of *nad9*. Only very weak evidence is found for recombination of nad2i709g2 across R180 into any of the three alternative arrangements B-D, B-E and B-F or for the reciprocal arrangement A-C (Fig. 5C).

Other than identical sequence repeats in distant locations, the mitogenome of *Haplopteris* carries copies of variably diverging sequences, sometimes closely spaced. Examples are an imperfect 39 bp IR embedding the *cox1* gene or an inverted sequence repeat of 1.1 Kbp embedding the *nad9* gene with (only) 70% sequence identity between the two copies. Direct repeats of an 85 bp sequence are located behind the *nad1* gene overlapped by a repeated tridecanucleotide motif (CCTCTACTGAGGG) at their ends.

The mitochondrial gene complement in *Haplopteris ensiformis*

Despite its highly complex structure, the *Haplopteris ensiformis* mitogenome has a rich gene complement (Tab. 1). All expected genes for subunits of the respiratory chain complexes I-V (*nad*, *sdh*, *cob*, *cox* and *atp* genes) are present, including *sdh3* and *sdh4* encoding subunits of complex II. Likewise, there is a surprisingly large set of genes for ribosomal proteins, given that many are absent even in the mtDNAs of more ancestral lineages. Notable is the retention of *rps7*, which is lacking in all available hornwort and lycopphyte mtDNAs but retained in *H. ensiformis*. Missing from the *Haplopteris* mtDNA is only the suite of *ccm* genes (*ccmB*, *ccmC*, *ccmF*) for cytochrome c maturation. Given their absence also in *Ophioglossum californicum* but their conservation in *Psilotum nudum* (Tab. 1) among the eusporangiate ferns, this evidently reflects a further independent loss of the *ccm* gene suite among ferns along with other phylogenetic deep losses in the lycophytes and hornworts.

The tRNA gene complement of the *H. ensiformis* mtDNA is particularly interesting owing to several tRNA genes of xenologous origin. Chloroplast-derived gene copies are present for trnF-GAA, trnMe-CAU, trnN-GUU, trnP-UGG and trnS-GGA (Tab. 1). Moreover, the *H. ensiformis* mtDNA also carries chlamydial-type tRNAs trnN-GUU (Fig. 4) and trnR-UCG (Fig. 3 and Fig. 4) described previously^{20,29}. This results in xenologous genes coexisting with their native counterparts for trnF-GAA and trnP-UGG. Genes for native tRNAs are lost for trnL-CAA, trnMe-CAU, trnN-GUU, trnR-ACG. The chlamydial-type trnR-UCG is remarkable since it exists in three slightly differing copies. Similarly notable is the presence of a trnL-UAG in *Haplopteris* that is absent in the eusporangiate ferns (Tab. 1).

RNA Editing in *Haplopteris ensiformis* mitochondria

Given the complex mitogenome structure, the parallel transcriptome and RNA editing analysis was fundamental to identify functional mitochondrial genes in the *Haplopteris ensiformis* mtDNA. All protein-coding genes were found to be affected by RNA editing. Altogether, we identified 1,618 events of mitochondrial RNA editing: 1,091 of the C-to-U type and 527 edits of the reverse U-to-C type (Suppl. Tab. 2). The abundance of RNA editing is highly biased among genes with the *cox1* mRNA being affected by 145 RNA editing sites in contrast to the *atp1* mRNA with only three edits, respectively.

While the very high coverage of RNA-Seq reads in chloroplasts allowed for determination of editing frequencies with high precision, the more than threefold abundance of edits in mitochondria allowed for a better statistical classification of edits categorized by their location and effect (Suppl. Tab. 2). Among the total of 1,618 mitochondrial edits, 1,171 introduce codon changes and of the latter more than 900 are strongly predicted and more than 100 others moderately or weakly expected. We here illustrate the prediction of editing sites by PREPACT with the example of *atp9* finding a perfect match between identified edits and expectations and for *atp6* with only minor deviations from the expected editing pattern (Suppl. Fig. 2). Notably, the strongly expected RNA editing events have an average editing frequency of 83%, much higher than only 53% on average for non-predicted changes of codon identities (Suppl. Tab. 2). Yet lower RNA editing efficiencies are observed for silent sites or the ones in non-coding regions, e.g. only 29% on average in 3'-UTRs (Suppl. Tab. 2). A notable exception

from efficient editing of coding regions is the *matR* maturase encoded in the terminal *nad1* intron (Suppl. Tab. 2). An intriguing observation was made for *matR*, the conserved maturase encoded in *nad1i728g2* (Tab. 1). Whereas identification of a start codon for this conserved, and only, mitochondrial group II intron maturase in flowering plants has been puzzling, we now find *matR* in *Haplopteris ensiformis* continuous with upstream *nad1* reading frame, accordingly, to be labelled *mat-nad1i728g2c* following a recent nomenclature proposal²⁸. Although only lowly edited, the numerous, and expected, events of RNA editing reconstituting conserved codons and including 15 stop codon removals confirm the functional role of *mat-nad1i728g2c* (Suppl. Tab. 2).

Particularly remarkable is that 44% of the reverse U-to-C edits (233 of 530) serve to re-convert stop codons into arginine or glutamine codons, an important issue to distinguish functional from dysfunctional pseudogenes. For example, within 20 codons upstream of *rps3i249g2*, six stop codons are removed from the *rps3* coding sequence, including one within the intron binding site (Suppl. Tab. 2).

A yet more dramatic example is the *rpl6-rps13-rps11* cluster with 20 genomic stop codons in a short region (Fig. 6). In fact, the *rps11* gene at DNA level initially appeared to be an amino-terminally truncated pseudogene but turned out to have a proper start codon created by C-to-U editing and a total of seven stops removed by reverse U-to-C editing within the first 20 codons of its reading frame. We wished to test how different RT-PCR-based “classic” approaches would perform in comparison to the RNA-Seq approach to detect RNA editing sites. Towards that end, we used three different strategies for cDNA synthesis using either random hexamers or specific primers targeting the 3'-end of *rps11* in edited or non-edited versions. Here, we made use of two edits in the 3'-UTR closely behind the *rps11eU466Q** stop codon editing. Sequencing of an internal amplicon revealed that many editing events were not confirmed in the cDNA sample primed with random hexamers with better performance by the specific primers and notably the one for the edited version of the 3'-end.

A striking bias concerns “silent” edits leaving codon identities unchanged. We observed only 32 silent U-to-C edits but the nine-fold amount (282) of silent C-to-U edits. Interestingly, silent C-to-U edits are frequently found to neighbor non-silent sites (“NESIs”) as has previously been observed for the huge editome in the *Selaginella uncinata* chloroplast³⁰.

Mitochondrial introns in *Haplopteris ensiformis* include a novel *trans*-splicing group I intron.

The *Haplopteris ensiformis* mitogenome shows notable differences to the intron complements in the two eusporangiate fern taxa (Tab. 1). Ancient group II introns nad1i477g2, nad1i669g2, nad5i1477g2 and nad7i917g2 are lost from the *H. ensiformis* mtDNA. *Vice versa*, group I intron cox1i395g1 and group II introns cox2i373g2 and rps14i114g2 in the *Haplopteris* mitogenome lack counterparts in *Ophioglossum* and *Psilotum* (Tab. 1). The group II introns and cox1i395g1 are evidently of ancient origin in the land plant lineage. The latter has previously been identified in liverworts but also in the leptosporangiate ferns and in the horsetail *Equisetum arvense*^{31,32}.

A striking example documenting the *Haplopteris ensiformis* mitogenome complexity concerns the ribosomal rRNA cluster with a peculiar arrangement featuring a disrupted *rrnL* gene (Fig. 3, Fig. 4). Maturation of the large ribosomal RNA requires *trans*-splicing of a broken group I intron, rrnLi825g1. Secondary structure modelling suggests base-pairing interaction of the two intron parts in the disrupted domain P9.0/P9.1 (Fig. 7). Intriguingly, rrnLi825g1 has a distant positional orthologue as a *cis*-splicing homologue in the charophyte alga *Chara vulgaris*³³. Despite overall only weak similarity, conserved regions include intron domains P7 and P8 known to contribute to the reactive core of group I introns. Three further introns are located in the downstream part of the *H. ensiformis* *rrnL* gene. Group II intron rrnLi833g2 is located only 8 bp downstream of the rrnLi825g1 3'-splice site and has distant homologues in liverwort mtDNAs. The two downstream group I introns rrnLi1897g1 and rrnLi1928g1 have hitherto only been identified serendipitously in *rrnL* gene samplings of other Polypodiales species. Notably, none of the four introns are present in the eusporangiate ferns, which feature continuous *rrnL* genes.

Laterally transferred chloroplast DNA fragments in the *Haplopteris* mitogenome

Among other issues, the assembly of the *Haplopteris ensiformis* mitogenome was much complicated by the fact that it contains ca. 80 inserts of chloroplast DNA of variable sizes and with variable degrees of sequence conservation. Similar to the repeat sequences, we annotated these cpDNA inserts with numbers indicating their sizes in base pairs preceded by 'cp' ([Suppl. Tab. 3](#)). While the separate chloroplast sequence inserts may have originated from fragmentation after insertion of large stretches of cpDNA, their different degree of sequence conservation rather argue for independent transfer events and likely document independent cpDNA insertions at different time points in evolution ([Fig. 8](#)). As an example, an array of seven likely independently acquired cpDNA inserts is present in the intergenic region between *nad5* and *sdh4* (Fig. 8A). This region includes cp4165, the largest continuous stretch of “promiscuous” cpDNA derived from the *ndhH-ndhA-ndhI-ndhG-ndhE* region, sharing 93 % sequence identity with the native chloroplast counterpart. This stretch is directly flanked by cp1271 derived from the chloroplast IR region encoding the *rrnL* gene and sharing even 99 % of identical nucleotides, likely indicating a yet more recent inter-organellar migration into the mitogenome. Chloroplast insert cp1039 downstream of *rps4* (Fig. 8B) is another example for a likely recently acquired lateral sequence transfer. In a phylogenetic analysis, it branches next to the homologous sequence from the *atpA-atpF* region of the now determined *Haplopteris ensiformis* cpDNA as sister to the counterpart in *H. elongata*, even despite the generally high sequence drift among Pteridaceae (Fig. 8B).

At the other end of the spectrum, some cpDNA inserts lack detectable homologous sequences in the *Haplopteris* chloroplast genomes and could only be recognized by similarities with the cpDNA of other taxa. Inserts cp364 and cp749 (Fig. 8A) are examples, which could only be identified by their similarity to sequences in the IR regions in the cpDNAs of genera like *Asplenium* or *Vittaria*. Yet more striking are the cases of morffos, the highly variable “mobile ORFs in fern organelles”²⁵. The chloroplast sequence insert cp2126 is an example along those lines (Fig. 8C), which includes a morffo element that

has a top sequence identity of 75% in the cpDNA of *Hemionitis subcordata* in the distant subfamily Cheilanthroideae.

Yet more striking is the case of morffo2, which is intact in the *Haplopteris elongata* cpDNA but truncated to its 5'-terminal 378 bp in the now determined *H. ensiformis* chloroplast genome (Fig. 1B). Intriguingly, chloroplast insert cp1712 in the mtDNA (Suppl. Tab. 3) includes significantly more of the 5'-end of the truncated morffo2 element (628 bp). Taken together, the slow sequence drift in the mitogenome offers examples allowing for a “molecular archeology” for former chloroplast DNA sequences that are not present any more in the recent chloroplast genome. In the latter case, it further supports the point for degeneration of the morffo elements in *H. ensiformis* rather than their independent origin in *H. elongata* (Fig. 1B).

Rickettsial DNA in the *Haplopteris ensiformis* mitogenome

The chlamydial-type *trnN-GUU* gene initially identified in the lycophyte *Phlegmariurus squarrosus*²⁹ and the chlamydial-type *trnR-UCG* gene found subsequently in the eusporangiate fern mitogenomes²⁰ are now also identified in the *Haplopteris ensiformis* mtDNA, further corroborating the concept of HGT from bacteria into the mitogenomes of early-branching vascular plants. Most strikingly, we now discovered numerous inserts of “Rickettsia-like” sequences in the *H. ensiformis* mitogenome. Altogether, we identified 30 variably degenerated protein coding regions evidently derived from rickettsial bacteria (see supplementary table 3). Similar to the labels for repeats and cpDNA inserts, we annotated the xenologous bacterial DNA inserts indicating their respective extension in base pairs, in this case preceded by an ‘x’ (Suppl. Tab. 3, [Fig. 9](#)).

We carefully verified the surprising observation of Rickettsia-like DNA inserts in the *Haplopteris ensiformis* mitogenome by PCRs anchoring in the flanking mtDNA regions and consistently corroborated the mitogenome assemblies, as we here exemplarily show for x625 representing a central coding region for the bacterial DNA recombination protein RmuC (Fig. 9A). Towards that end we used independent DNA preparations from two *H. ensiformis* isolates from separate locations and,

vice versa, included material from *Vittaria lineata*, a closely related species that grows next to one of the *H. ensiformis* populations in the Botanic Garden Bonn. PCR products were consistently obtained for the two independent *H. ensiformis* isolates, but not for the *V. lineata* sample (Fig. 9A). Sequencing of the PCR products fully confirmed the mitogenome assembly, interspersed by Rickettsial-like DNA insertions. Moreover, none of the bacterial insert sequences pointed to contamination by living bacteria as they mostly revealed characteristic degeneration of the protein coding genes to pseudogenes. Finally, we observed continuities of the average read coverages continuing from flanking mitochondrial sequences into the bacterial DNA inserts (Fig. 9). We do not assume functional expression of the xenologous bacterial genes given that we could detect only negligible RNA coverage for some of the regions carrying bacterial DNA inserts.

After insertion into the *Haplopteris* mtDNA, the Rickettsia-type coding sequences degenerated by recombination and sequence drift. Some protein sequence similarities, however, remain astonishingly high, likely indicating very recent horizontal gene transfers (Suppl. Fig. 3 and Suppl. Fig. 4). Specifically, top similarities were often observed with the corresponding sequences from *Caedimonas varicaedens*³⁴, followed by slightly lower similarities with homologous loci in *Cand. Paracaedimonas acanthamoebae*, other *Caedibacter spp.* or *Cand. Nucleicultrix amoebiphila*. These species belong to the family Holosporaceae among Rickettsiales *sensu lato*. The family Holosporaceae is alternatively considered to be a separate order of its own, the Holosporales.

As in the case of the chloroplast sequence inserts, the variable degrees of sequence degeneration suggest independent events of horizontal transfers at different time points in evolution. Alternatively, the insertion of larger xenologous genomic regions followed by later fragmentation in the mitogenome followed by different degrees of sequence degeneration may be possible. One evident example is the bacterial *HscA-RmuC* region (Fig. 9B) located between the *nad1* gene and cp232 (Fig. 9B, see Suppl. Tab. 3). The *HscA* coding region is disrupted by an insert in the coding sequence and the downstream *RmuC* gene is truncated with its central region located as x625 between the *cox2* gene and the *rrnL-rrnS* gene cluster, directly flanking cpDNA insert cp416 (Fig. 9A). Despite these discontinuities, the

degree of amino acid sequence conservation reaching up to 94 % identity is astonishing and phylogenetic analysis allows a close affiliation of the xenologous RmuC region in the *Haplopteris* mitogenome with *Caedimonas varicaedens* among the Holosporales (Fig. 9). Another xenologous insert x888 (Fig. 9B) carrying parts of the coding region for OMBB, an outer membrane beta-barrel domain containing protein, is much more degenerated and has significantly lower similarities to top-scoring hits with less clearly defined Rickettsiales bacteria (33% identity, 50% similarity).

Whereas most of the xenologous bacterial DNA inserts show fragmentation and considerable degeneration and of protein-coding regions, there are also striking counterexamples like x1850, an insert of ca. 1.8 Kbp, containing the *serS-surE-nlpD* region located embedded between cp763 and cp1456 (Suppl. Tab. 3). The coding region of the upstream serine tRNA ligase SerS is N-terminally truncated and carries two stop codons but the reading frame of the downstream coding region for the 5'/3'-nucleotidase SurE is perfectly conserved and shares even 97 % sequence similarity with its *Caedimonas varicaedens* counterpart (Suppl. Fig. 3). In this case a similarity of 81% can even be identified at the nucleotide sequence level (Suppl. Fig. 4), clearly suggesting the bacterial donor to be closely related to *C. varicaedens*. Intriguingly, the end of this bacterial sequence insert contributes to repeat R295.

Examples for particularly large stretches of Holosporaceae DNA inserts include x2170 derived from the bacterial XpsD-region located behind the mitochondrial *nad4L* gene (Suppl. Tab. 3) and x1920, originating from the bacterial MfdD-RimM-TrmD-RpL19 region inserted upstream of *rps4* (Fig. 9C, Suppl. Tab. 3). In the latter case, the gene for MfdD is severely 5'-truncated, the gene for RimM appears intact, the TrmD reading frame is degenerated by a stop codon and the RpL19 sequence is 3' truncated (Fig. 9C).

In some cases, top sequence similarities clearly identify an origin of HGT sequences from Rickettsiales, but not necessarily from within the Holosporaceae. Most notable is an assembly of sequence inserts of altogether more than 7 Kbp downstream of the *atp1* gene (Suppl. Tab. 3) that

carries the suite of genes encoding *murA* and two CoA-carboxylase subunits in opposite direction, which are nearly unaffected by sequence degeneration except for in-frame stop codons (Suppl. Fig. 3). At nucleotide level, a top similarity of 74 % identical nucleotides is observed with a not further identified Rickettsiaceae bacterium isolate PMG_002 (Suppl. Fig. 4). Intriguingly, the *VirB8/B9* coding sequences of two P-type conjugative transfer proteins are located ca. 1.2 Kbp downstream of the *murA* homology. Top nucleotide sequence similarities are again observed for Rickettsiaceae bacterium isolate PMG_002. However, the *virB8/B9* and the *MurA* coding sequences are not connected in that genome, indicating either a related and yet unidentified Rickettsiales donor, separate gene transfers or subsequent rearrangements after copy-transfer into the *Haplopteris* mtDNA.

Finally, two large stretches of protein coding regions are located upstream of *nad9*, running in opposite direction (labelled Bact-ORF 1 and 2, respectively), but cannot be assigned taxonomically owing to much lower similarities of only 40-50% at protein sequence level: A hypothetical ORF, possibly RNA polymerase, highest similarity with a *Magnetovibrio* sp. database entry (MBM08139) and a DNA-polymerase, highest similarity with a *Zoogloea* sp. sequence entry (KAB2964018).

Discussion

A first leptosporangiate fern mitogenome assembly

The large Pteridaceae family contains about 10 % of extant ferns species^{35–37}. Among them, the Adiantoid and Vittarioid sub-families (see Fig. 8) show particularly high levels of substitution rate heterogeneity^{38,39}. We became particularly interested in the taxon given the apparently dynamic evolution of mitochondrial introns and RNA editing²². Likewise, a striking diversity of chloroplast RNA editing has been reported in the genus *Adiantum*⁴⁰.

The genomes of the two endosymbiotic organelles in *Haplopteris ensiformis* here reported are prime examples highlighting the discrepancy between the conservative evolution of chloroplast DNA and the highly dynamic evolution of vascular plant mitochondrial genomes. Among the recently reported flowering plant mitogenomes, the one of the holoparasite *Ombrophytum subterraneum*⁴¹ is

a case in point documenting not only a multi-chromosomal structure but multiple evidence for HGT from its host plants. Like in other cases of multi-chromosomal plant mitogenomes, e. g. the hundreds of different mitochondrial chromosomes in some *Silene* species, no efforts are made any more to come up with the display of a hypothetical, and likely misleading “master-circle”^{42,43}.

As a first representative for the large clade of leptosporangiate ferns with ca. 10,000 species, the *H. ensiformis* mitogenome again adds to the list of astonishing molecular peculiarities that are hallmarks of vascular plant mtDNAs^{44,45}. Given its extraordinary complexity, it comes as no surprise that no other leptosporangiate mitogenome has previously been assembled despite multiple NGS efforts including the water fern genera *Azolla* and *Salvinia*^{46,47} or the “flying spider-monkey” tree fern *Alsophila spinulosa*⁴⁸ or, most recently, the model ferns *Adiantum capillus-veneris*⁴⁹ and *Ceratopteris richardii*⁵⁰. Likewise no mitogenome was reported in a genome assembly effort for the lycophyte *Isoetes taiwanensis*⁵¹, possibly owing to even higher complexity than the ones reported previously for *Isoetes engelmannii* and *Selaginella moellendorffii*^{7,8}.

We hence speculate that highly complex mitogenomes like the one reported here may be a general feature of leptosporangiate ferns also outside of the Polypodiales. Intriguingly though, and despite its highly dynamic structural evolution, the *H. ensiformis* mtDNA contains a surprisingly rich set of “classic” mitochondrial genes (Tab. 1) when compared to the gene complement of other taxa including the eusporangiate ferns⁵².

Mitochondrial intron dynamics in ferns

More notable than the gene complements are the diverging mitochondrial intron complements now identified in *Haplopteris ensiformis* in comparison to the previously analyzed eusporangiate ferns, extending earlier conclusions that much more intron dynamics is present in monilophytes in comparison to their seed plant sister clade^{21,22,32,52}. The *Haplopteris* mitogenome reveals retention of evidently ancient introns that have been lost in the eusporangiate ferns like *cox1i395g1*, *cox2i373g2*, *rps14i114g2* and *rrnLi833g2* (Tab. 1).

Surprisingly though, the heavily recombining mtDNA of *Haplopteris ensiformis* has not resulted in disrupted group II introns like in angiosperms or in gymnosperms where ever more transitions to *trans*-splicing have been observed recently⁵³. In contrast, most group II introns (11 of 15) found to be *trans*-splicing in at least one seed plant lineage (cox2i373g2, nad1i394g2, nad1i728g2, nad2i542g2, nad2i1282g2, nad4i461g2, nad4i976g2, nad4i1399g2, nad5i1455g2, nad7i209g2 and rpl2i846g2) are present in conventional *cis*-arrangements in the *Haplopteris* mtDNA fully in line with the early evolutionary conclusion that *trans*-splicing introns in seed plants originate from *cis*-arranged ancestors in early-branching plant lineages^{54,55}. We speculate that transitions from *cis*- to *trans*-splicing group II introns may rely on co-evolving protein splicing factors that are present in the seed plant but not in the monilophyte lineage. The remaining four group II introns known to exist in a *trans*-splicing state in at least some spermatophytes (cox2i691g2, nad1i669g2, nad5i1477g2 and nad7i917g2) have been lost altogether from the *H. ensiformis* mitogenome (Tab. 1).

In the light of the above it is all the more surprising to find rrnLi825g1 as a “novel” group I intron of yet unclear ancestry in a *trans*-splicing arrangement in the idiosyncratic *rrnL* gene makeup in the *H. ensiformis* mitogenome. In contrast to the numerous examples of *trans*-splicing group II introns in plant organelles alone, *trans*-splicing group I introns appear to exist much more rarely in nature. First reports of *trans*-splicing group I introns in the *Isoetes engelmannii* mitogenome⁷ and in the mtDNA of *Trichoplax*⁵⁶ have been followed by recognition of *trans*-splicing group I intron cox1i744g1 in *Helicosporidium* mtDNA⁵⁷ and of two *trans*-spliced group I introns in *Gigaspora margarita* mtDNA⁵⁸. Remarkably, the *trans*-splicing group I intron cox1i395g1 in the *Isoetes engelmannii* mtDNA⁷ exists in a conventional, *cis*-arranged version in the *H. ensiformis* mitogenome.

The mitochondrial *rrnL* gene is entirely devoid of introns not only in seed plants and lycophytes but also in hornworts²⁸ and even in the eusporangiate ferns²⁰ whereas it features four introns in the *Haplopteris ensiformis* mtDNA: rrnLi825g1, rrnLi833g2, rrnLi1897g1 and rrnLi1928g1 (Fig. 7). The latter two group I introns have already been documented serendipitously in sequence samplings covering parts of the mitochondrial *rrnL* gene⁵⁹ and are conserved also outside of Pteridaceae in taxa as least as

distant as tree ferns (Cyatheaales, e.g. *Plagiogyria stenoptera*, accession DQ647877). The origins of these introns remain unclear as they neither share sequence similarities anywhere else in the plant lineage nor among fungi as is the case for the “rampant invader” group I intron *cox1i726g1* sporadically occurring in angiosperms^{60,61}.

Verifying native and foreign sequences in the organelle genomes

The examples of the *Haplopteris ensiformis* organelle genomes reported here document that parallel analysis of transcriptome along with genome NGS sequencing data is essential to characterize genes as functional or dysfunctional owing to the complex maturation processes including C-to-U and U-to-C RNA editing and *cis*-splicing or *trans*-splicing of split genes and introns. Moreover, the highly complex mitogenome of *H. ensiformis* is a prime case showing that very careful investigations of “alien” sequences in NGS data to tell them apart from native chloroplast DNA or bacterial contaminations may be needed in such complex cases of organelle genome structures. Aside from our experimental verifications, we note that parts of rickettsial DNA inserts in *H. ensiformis* also seem to be present in database entries reporting partial mtDNA sequences of the ferns *Asplenium nidus* (FR669448) and *Dryopteris crassirhizoma* (MW732172). Moreover, we suggest the careful re-evaluation of some fern cpDNA entries (like *Dipteris conjugata* KP136829, *Polypodium vulgare* MT984517, *Cystopteris protrusa* KP136830 or *Selliguea yakushimensis* MN623352) that seem to contain mtDNA stretches as possible artefacts. Certainly, however, once verified and when the likely similarly complex mitogenomes of those taxa would be assembled in the future they may document interesting inter-organellar DNA transfer from mitochondria to chloroplasts.

Lateral sequence transfers: cpDNA insertions in the *Haplopteris* mitogenome

The first evidence for lateral transfer of “promiscuous” chloroplast DNA into a plant mitogenome has already been documented 40 years ago⁶² and thereafter found to be merely a standard feature in many seed plant mtDNAs. Most interestingly, among the numerous cpDNA insertions in the mtDNA of *Haplopteris ensiformis* we now find examples evidently documenting ancient cpDNA features that even are not present any more in the recent plastomes, as here seen for the mysterious morffo

elements. The origin and dynamics of these only recently described “mobile ORFs in fern organelles”²⁵ is presently still enigmatic. Here, we find that apparently intact morffos in the cpDNA of *Haplopteris elongata* have evidently disintegrated in the plastome of *H. ensiformis* (Fig. 1B) but that their counterparts and evidence for yet other morffos is present in its mitogenome as “evolution’s misplaced witnesses”.

Origins of bacterial sequences in the *Haplopteris* mtDNA

The role of Horizontal Gene Transfers (HGT) is increasingly appreciated not only for bacterial evolution but also in the evolution of eukaryotic genomes including ferns^{63,64}. In several cases, host-parasite interactions are key to the events of HGT^{41,65} and such interactions in nature may also be responsible for gene transfer into fern mtDNA⁶⁶. In particular, after first reports on horizontal plant-to-plant transfer of mtDNAs^{67,68}, it is meantime well understood that HGT has contributed to many seed plant mitogenomes⁶⁹. The most outstanding example is the case of the early-branching flowering plant *Amborella trichopoda* having integrated into its mitogenome not only numerous stretches of mtDNA from other angiosperms but also the near-complete mitogenomes of two mosses^{19,70}. A new dimension of HGT into plant mitogenomes opened up with the discovery of tRNA genes from chlamydial origins into very early tracheophyte lineages²⁹. Yet more recently, it was found that sequences of fungal origin have been horizontally transferred early into mitogenomes of the Orchid family⁷¹.

Here, we now report on a multitude of Rickettsia-like genome insertions as one of the most astonishing findings emerging from the assembly and analysis of the *Haplopteris ensiformis* mitogenome. Intriguingly, Rickettsiales are known to be obligate intracellular parasites and at the same time assumed to be the extant alpha-proteobacterial lineage most closely related to the progenitor of the eukaryotic mitochondrion. The bacterial DNA inserts in the *H. ensiformis* mtDNA are most likely derived from species closely related to (Cand.) *Caedibacter acanthamoebae*³⁴ of the Holosporales (or, alternatively, Holosporaceae among Rickettsiales). *Caedibacter* (or *Caedimonas*) endosymbionts transfer the “killer trait” to their *Paramecium* hosts^{72,73}. Other Rickettsia are known to be associated

with arthropods, leeches and protists and Rickettsia-like organisms (RLOs) and are associated not only with human or animal diseases but also with numerous plant diseases, for example the Rickettsia endosymbiont of the tobacco whitefly *Bemisia tabaci*. Moreover, Rickettsia have also been associated with a papaya disease⁷⁴ and have been identified eustigmatophyte algae⁷⁵ and in the green alga *Mesostigma viride*⁷⁶. It will be very interesting to see whether Rickettsial or related bacterial DNA insertions will also be identified in further leptosporangiate mitogenomes and to ultimately identify the exact donor species and the biological mechanisms of the HGT processes. Notably, the xenologous bacterial DNA regions identified in the *H. ensiformis* mitogenome include similarities to the IS481 family transposase and virB8 and virB9 homologs (Suppl. Tab. 3). The latter genes are commonly found among mobile IS elements and are associated with conjugative gene transfer amongst Rickettsia species⁷⁷.

Methods

Plant material

The Bonn University Botanic Garden kindly provided plant material for *Haplopteris ensiformis* (xx-0-BONN-24687) and *Vittaria lineata* (xx-0-BONN-17295). Species identities were independently verified by PCR amplification and sequencing of the *rbcl* and *atpA* locus revealing complete sequence identities with independent sequence accessions (*Haplopteris ensiformis* KX164999 and MH359250 and *Vittaria lineata* EF473712 and KU744782).

Organelle genome sequencing and assembly

The Qiagen DNeasy Plant Mini Kit was used for DNA isolation and the Sigma-Aldrich plant RNA isolation Kit for RNA preparation followed by ribosomal RNA depletion using the RiboMinus Plant Kit for RNA-Seq (Thermo Fisher Scientific). RNA quality was checked with a Qubit Fluorometer for RIN-values of at least 0.9. Genome and transcriptome sequencing (paired-end whole genome sequencing and RNAseq) was done commercially at the BGI on an Illumina platform. Raw read data were evaluated with FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). No adapter artifacts could be detected. The MEGAHIT software^{78,79} was used for *de novo* whole genome assembly with independent

runs using three different settings (“strict”, “default” and “relaxed”). The “strict” assembly was run with the minimum k-mer option available. The third assembly was run with parameters set for a high ultra-complex metagenomics dataset). Complete assembly of cpDNA was performed with NOVOPlasty⁸⁰ using contigs from the MEGAHIT assembly with conserved chloroplast genes as seeds. RNA reads were assembled with the Trinity software^{81,82}. The BLAST⁸³ suite was used to initially identify contigs with chloroplast or mitochondrial gene content using available lycophyte and fern organelle genomes. Raw read data and assembly data have been deposited under BioProject accession no. PRJNA862965. Whole genome assembly contigs could be clustered into two sets based on k-mer coverage (mt/cp 1:10). After connections of contigs were verified by PCRs (see below), contigs of the mitochondrial genome were connected by hand.

Verification of mtDNA arrangements

PCRs were used to independently verify the highly complex arrangements of the *Haplopteris ensiformis* mitochondrial DNA resulting from multiple repetitive sequences and insertions of chloroplast and xenologous DNA. PCR amplicons were designed with primers anchoring in neighboring sequence regions of evident mitochondrial identity, preferably coding regions. Special care was taken to investigate repeated sequence for potential recombination creating alternative arrangements of flanking sequences. To best avoid false positives suggesting active recombination resulting from artificial template switches we used a strategy of template-switch avoiding “tsa”-PCRs. To that end, a mix of gel-eluted PCR fragments containing a repeat sequence in different sequence environments (AB and CD) was used to obtain products reflecting a reciprocal exchange of flanking regions (AD and CB). A series of template dilutions (1:10, 1:20, 1:30, 1:40, 1:50 and 1:60) and numbers of PCR cycles (15, 20, 25 and 30) were tested and adjusted to determine the threshold for artificial production of template-switch products.

Identification of mitochondrial genes avoided routine pipelines but started from homologues in the mtDNA of diverse taxa, including the liverwort *Marchantia polymorpha*, the lycophyte *Phlegmariurus*

615 *squarrosus* and the gymnosperm *Ginkgo biloba*. Gene identities were verified by the parallel
616 transcriptome studies to confirm intron splicing and C-to-U and U-to-C RNA editing. Identification of
617 tRNA genes combined the use of tRNAscan-SE⁸⁴ and sensitive BLASTN searches using a tRNA query set
618 including the recently identified chlamydial tRNA xenologues in early tracheophytes^{20,29}. To identify
619 DNA similarities including repeats, chloroplast DNA or xenologous bacterial DNA insertions, we used
620 sensitive BLASTN or XBLAST similarity searches (word sizes = 7 or 3, respectively) and strict random
621 expectancy threshold cutoffs of 1e-10. On nucleotide level, this translates approximately into
622 identification of identical repeated sequences of ca. 40 bp (i.e. slightly larger than the conserved
623 domain V of group II introns) or respective larger, but less similar, regions.

624 Transcriptome studies and determination of RNA editing sites

625 Transcriptome studies were used to determine all intron splicing sites. The identification of RNA editing
626 sites was done as previously described²⁸. Briefly, DNA and RNA reads were mapped against the
627 organelle contig sequences using GSNAP⁸⁵ and JACUSA⁸⁶ was used to determine RNA-DNA differences.
628 Thresholds were set to minimally 30 reads and RNA editing efficiencies of at least 1% for chloroplast
629 and at least 5% for mitochondrial transcripts for a strict determination of C-to-U and U-to-C RNA
630 editing sites, respectively. Selected loci were analyzed independently by RT-PCR-based cDNA analyses
631 and sequencing for various reasons like unexpectedly inefficient RNA editing at certain sites or because
632 of coexisting pseudogene copies as discussed under results (chloroplast *rpoC1* and mitochondrial
633 genes *atp1*, *atp8*, *nad5*, *nad7*, *rrnL* and the *rpl6-rps13-rps11* co-transcript).

Data availability and Sequence accessions

Haplopteris ensiformis primary nucleotide sequence reads are submitted to the sequence read archive (SRA) under BioProject accession number PRJNA862965. The assembled chloroplast genome is deposited under accession number OM867544 and the assembled mitogenome chromosomes are available under accession numbers OM867545 to OM867553.

Acknowledgements

The authors wish to thank the Molecular Evolution group for stimulating discussion and productive interaction, especially Philipp Gerke for his contributions to establish pipelines for the analysis of RNA editing in NGS data. The authors gratefully acknowledge the support for this project by computing time provided by the Paderborn Center for Parallel Computing (PC²).

Author contributions

SZ did wet lab work and established bioinformatic pipelines. MP helped with nucleic acid preparations and molecular cloning. SZ and VK analyzed data and prepared figures. VK wrote the manuscript and all authors edited and approved the final manuscript version.

648 Table 1. The mitochondrial gene and intron complement of *Haplopteris ensiformis*.

Gene	intron	<i>Psilotum nudum</i>	<i>Ophioglossum californicum</i>	<i>Haplopteris ensiformis</i>	Gene	intron	<i>Psilotum nudum</i>	<i>Ophioglossum californicum</i>	<i>Haplopteris ensiformis</i>
<i>atp1</i>		+	+	+ ^a	<i>rps1</i>		+	+	+
<i>atp4</i>		+	+	+	<i>rps1i25g2</i>		+	0	+
<i>atp6</i>		+	+	+	<i>rps2</i>		+	+	+
<i>atp8</i>		+	+	+ ^a	<i>rps3</i>		+	+	+
<i>atp9</i>		+	+	+	<i>rps3i74g2</i>		+	+	+
<i>ccmB</i>		+	0	0	<i>rps3i249g2</i>		+	0	+
<i>ccmC</i>		+	0	0	<i>rps4</i>		+	+	+
<i>ccmF</i>		+	0	0	<i>rps7</i>		+	+	+
<i>ccmFCi829g2</i>		+	0	0	<i>rps10</i>		+	+	+ ^a
<i>cob</i>		+	+	+ ^a	<i>rps11</i>		+	+	+
<i>cox1</i>		+	+	+	<i>rps12</i>		+	+	+
<i>cox1i395g1</i>		0	0	+	<i>rps13</i>		+	+	+
<i>cox1i624g2</i>		+	0	0	<i>rps14</i>		+	+	+
<i>cox2</i>		+	+	+	<i>rps14i114g2</i>		0	0	+
<i>cox2i373g2</i>		0	0	+	<i>rps19</i>		+	+	+
<i>cox3</i>		+	+	+	<i>rrn5</i>		+	+	+
<i>matR</i>		+	+	+	<i>rrnL</i>		+	+	+
<i>nad1</i>		+	+	+	<i>rrnLi825g1</i>		0	0	T
<i>nad1i258g2</i>		+	+	+	<i>rrnLi833g2</i>		0	0	+
<i>nad1i394g2</i>		+	0	+	<i>rrnLi1897g1</i>		0	0	+
<i>nad1i477g2</i>		+	+	0	<i>rrnLi1928g1</i>		0	0	+
<i>nad1i669g2</i>		+	+	0	<i>rrnS</i>		+	+	+
<i>nad1i728g2</i>		+	+	+	<i>sdh3</i>		+	+	+
<i>nad2</i>		+	+	+	<i>sdh4</i>		+	+	+
<i>nad2i156g2</i>		+	+	+	<i>tatC</i>		+	+	+ ^a
<i>nad2i542g2</i>		+	+	+	<i>trnA-ugc</i>		+	+	+ ^b
<i>nad2i709g2</i>		+	+	+	<i>trnC-gca</i>		+	+	+
<i>nad2i1282g2</i>		+	+	+	<i>trnD-guc</i>		+	+	+
<i>nad3</i>		+	+	+	<i>trnE-uuc</i>		+	+	+
<i>nad4</i>		+	+	+	<i>trnF-gaa</i>		+	+	+
<i>nad4i461g2</i>		+	+	+	<i>trnF-gaa cp</i>		0	0	+ ^c
<i>nad4i976g2</i>		+	+	+	<i>trnG-gcc</i>		+	+	+
<i>nad4i1399g2</i>		+	+	+	<i>trnG-ucc</i>		+	+	+
<i>nad4L</i>		+	+	+	<i>trnH-gug</i>		+	+	+ ^b
<i>nad5</i>		+	+	+	<i>trnI-cau</i>		+	+	+
<i>nad5i230g2</i>		+	+	+	<i>trnK-uuu</i>		+	+	+
<i>nad5i1242g2</i>		+	0	+	<i>trnL-caa</i>		+	-	+
<i>nad5i1455g2</i>		+	+	+	<i>trnL-uaa^{b,c}</i>		+	+	+
<i>nad5i1477g2</i>		+	+	0	<i>trnL-uag</i>		0	0	+
<i>nad5i1872g2</i>		+	+	+	<i>trnMe-cau cp</i>		+	+	+ ^c
<i>nad6</i>		+	+	+	<i>trnMi-cau</i>		+	Ψ	+
<i>nad7</i>		+	+	+	<i>trnN-guu cm</i>		0	0	+
<i>nad7i140g2</i>		+	+	+	<i>trnN-guu cp</i>		+	+	+ ^c
<i>nad7i209g2</i>		+	0	+	<i>trnP-ugg</i>		+	+	+
<i>nad7i676g2</i>		+	+	+	<i>trnP-ugg cp</i>		0	0	+ ^c
<i>nad7i917g2</i>		+	+	0	<i>trnQ-uug</i>		+	+	+
<i>nad9</i>		+	+	+	<i>trnR-acg</i>		+	+	0
<i>rpl2</i>		+	+	+	<i>trnR-ucg cm</i>		+	+	+ ^b
<i>rpl2i846g2</i>		+	+	+	<i>trnR-ucu</i>		+	+	+
<i>rpl5</i>		+	+	+	<i>trnS-gcu</i>		+	+	+
<i>rpl6</i>		+	+	+	<i>trnS-gcu cp</i>		0	0	+ ^c
<i>rpl16</i>		+	+	+	<i>trnS-gga cp</i>		0	0	+
					<i>trnS-uga</i>		0	+	+
					<i>trnV-uac</i>		0	+	+
					<i>trnW-cca</i>		+	+	+
					<i>trnY-gua</i>		+	+	+ ^b

649 **Table 1.** List of mitochondrial genes and group I (g1) and group II (g2) introns in the *Haplopteris*
650 *ensiformis* mitogenome in comparison to the ones of the eusporangiate ferns *Psilotum nudum* and
651 *Ophioglossum californicum*. The added ‘cm’ or ‘cp’ indicate chloroplast-derived or chlamydial-type
652 tRNA genes, respectively. Superscript addendums: ‘a’ indicates co-existing large pseudogene copies,
653 ‘b’ indicates co-existing functional copies with only minor sequence differences, ‘c’ indicates that

654 chloroplast tRNA genes are part of extended cpDNA inserts. Features distinguishing the *Haplopteris*
655 *ensiformis* mtDNA from both eusporangiate ferns are highlighted in red, i.e. the presence of a
656 chlamydial *trnN-guu* gene, of chloroplast-derived *trnS-gcu* and *trnS-gga* genes, of introns *cox1i395g1*,
657 *rps14i114g2* and of four introns in the *rrnL* gene including the *trans*-splicing intron *rrnLi825g1* vs. the
658 absence of a *trnR-acg* gene and four other group II introns. The maturase “*matR*” in the terminal *nad1*
659 intron, systematically labelled *mat-nad1i728g2c*, is in-frame with the upstream *nad1* coding region in
660 *H. ensiformis*.

Figure Legends

Figure 1. [The *Haplopteris ensiformis* cpDNA.](#)

A. *Haplopteris ensiformis* reveals a typical plant circular plastome structure consisting of a large (LSC) and a small (SSC) single-copy region separated by a pair of inverted repeats (IR) and an expectedly conserved, ancestral gene and intron complement. The genome map was created using OGDRAW⁸⁷. Gene categories are indicated in the legend. Numbers in parentheses indicate the amount of C-to-U (blue) and U-to-C (red) RNA editing for the respective genes. Creations of start or stop codons by C-to-U editing are indicated by symbols ‘>’ and ‘*’ and the removal of stop codons by U-to-C editing is indicated by the exclamation marks, respectively. **B.** The cpDNA of *Haplopteris elongata* (accession MH173086) features two “morffo” elements (“mobile ORFs in fern organelles”) in the IR region between *rrn5* and the 3’-part of the *trans*-splicing *rps12* gene. Recognizable sequence homologues of morffo2 can presently only be identified in *Cyclosorus interruptus* (accession MN599066, Thelypteridaceae) and *Histiopteris incisa* (accession MH319942, Dennstaedtiaceae) and a homologue of morffo1 (orange) can presently only be found in the distant fern *Hymenophyllum holochilum* (accession MH265124, Hymenophyllales). Only the upstream part of morffo2 (378 bp) is present in the *H. ensiformis* plastome, while a cpDNA insert in its mitogenome contains an extended region of 628 bp.

Figure 2. [The *accD* gene example for chloroplast RNA editing in *Haplopteris ensiformis*.](#)

A. Sequence alignment of the *H. ensiformis accD* gene below its homologue in *Psilotum nudum* shown as one selected example reference out of 22 used for prediction of RNA editing. Alignment was created by PREPACT⁸⁸ with identical nucleotides and amino acid shown in grey font and predicted C-to-U RNA editing in blue and reverse U-to-C editing in red. Codons framed by boxes were confirmed as editing sites. The stippled rectangle highlights potential reverse edit accDeC580FL remaining unconfirmed but strongly suggested by the chloroplast editome references in PREPACT 3.0 with the exception of *Ginkgo biloba*. The remaining cases are weak predictions only that are not supported by the majority of other editome references. **B.** The list of expected and observed *accD* edits including those in the 5’ and 3’-UTRs and the respective editing frequencies observed with additional remarks.

Figure 3. [The *Haplopteris ensiformis* mtDNA: contigs A-C and chromosome 5](#)

Haplopteris ensiformis mtDNA contigs A (hatched), B (grey) and C (dotted) can be connected into a circular chromosome of 217,669 bp. Recombination breakpoints are numbered for each contig and preceded by a small ‘r’ with radial lines in the circular map indicating transitions into other contigs, allowing for numerous alternative mtDNA arrangements. Eight further circular chromosomes, as listed on top, connect sequences of chromosome V with nine further mtDNA contigs D-L as shown in [figure](#)

4. Contig gene maps were created using the SnapGene Viewer software. Native mitochondrial protein coding and tRNA gene sequence are given in lighter and darker blue, respectively, and introns are indicated with black arrows. Numbers next to genes indicate C-to-U (blue) and U-to-C RNA edits (red) with additional symbols indicating removal (!) or creation of stop (*) or start (>) codons, respectively. Introns are indicated with stippled lines and additional black arrows and their standardized labels. Ribosomal RNA genes (here on contig A) are shown in red with the *rrnL* gene featuring a complex gene structure requiring *trans*-splicing via the disrupted group I intron *rrnLi825g1* and PSX labels indicate pseudogene fragments shown in grey. The peculiar case of the *trans*-spliced group I intron *rrnLi825g1* is highlighted in yellow. Genes for tRNAs of chloroplast or bacterial origin are indicated in green or purple, respectively. For clarity, no other chloroplast or bacterial DNA insertions are shown here. The latter are listed together with annotated features in supplementary table 3.

Figure 4. [The *Haplopteris ensiformis* mtDNA: contigs D-L and chromosomes 1-4 and 6-9.](#)

The *Haplopteris ensiformis* mtDNA contigs D-L are linked to recombination points in chromosome 5 (Fig. 3) or within themselves, creating further and/or alternative mtDNA arrangements. Recombination endpoints are labeled and the display of contigs and labels for genes and RNA editing events is as in figure 3. Possible, circular chromosomal structures chr1 to chr4 and chr6 to chr9 are shown.

Figure 5. [Repeats and recombination in the *Haplopteris ensiformis* mitogenome.](#)

Recombinations across repeats (in orange) R596 (A), repeats R203 and R 513 (B) and R180 (C) was investigated by tsa-PCR ("template-switch-avoiding") strategies. **A.** R596 is identically present in domains IV of group II introns *nad5i1242g2* and *rrnLi833g2*. Average read coverage of the flanking single copy regions were ca. 150 x for *nad5* (arrangement A-C) and ca. 350 x for *rrnL* (arrangement B-D), apparently adding up to ca. 500 x for R596. PCR products are obtained for the expected gene continuities (AC, BD) with only minor evidence for reciprocal exchanges (AD, BC). **B.** All combinations of flanking sequences (AC-BD) are identified for recombination across R203 whereas a clear bias is seen for recombination across R513 where a product for primer combination B-C remains undetected. **C.** One copy of repeat R180 is located in intron *nad2i709g2*, another one downstream of the *nad9* gene. The region between R180 and *nad9* contains additional repeats R137 and R259 and all consecutive recombination products can be found whereas there is only very weak evidence for recombination across the R180 copy located in *nad2i709g2*.

Figure 6. [Mitochondrial RNA editing in *Haplopteris ensiformis*: the *rpl6-rps13-rps11* case.](#)

The alignment exemplarily shows RNA editing heavily affecting the *rpl6-rps13-rps11* co-transcript including the removal of seven stop codons within the first 20 codons of the *rps11* reading frame. An internal PCR amplicon covers 46 editing sites from *rpl6eC106**R to *rps11eC136**Q. Synthesis of cDNA

was primed either with random hexamers (n6) or with specific primers covering the end of *rps11* including the stop codon generation and two edits in the 3'-UTR in an edited (P+) or unedited (P-) version. The alignment displays edits (C-to-U in blue and U-to-C in red) clearly revealed in the RT-PCRs primed by the three different approaches and the results from the RNA-Seq data (RRM). Silent codon edits shown below the protein alignment were exclusively identified in the latter.

Figure 7. [A trans-splicing group I intron in the *Haplopteris ensiformis* mitochondrial *rrnL* gene.](#)

A. Maturation of *rrnL* includes *trans*-splicing of the disrupted group I intron *rrnLi825g1*. Three additional introns are removed from the downstream part of *rrnL*. Group II intron *rrnLi833g2* has a homologue conserved in the mtDNAs of liverworts and the downstream group I introns *rrnLi1897g1* and *rrnLi1928g1* are conserved in other Polypodiales species. Splice sites of *rrnLi825g1* and *rrnLi833g2* (black and grey triangles, respectively) frame the tiny second *rrnL* exon of only eight nucleotides. **B.** Secondary structure model of the disrupted, *trans*-splicing group I intron *rrnLi825g1*. The figure was generated using the VARNA software ⁸⁹. Regions for base-pairing between the upstream and downstream parts of *rrnLi825g1* are found in the group I intron secondary structure P9.1. Intron *rrnLi825g1* intron has a positional orthologue in the sweet-water alga *Chara vulgaris*. Green and yellow shading indicates identical nucleotides and transitions, respectively.

Figure 8. [A multitude of chloroplast DNA inserts in the *Haplopteris ensiformis* mitogenome.](#)

Selected examples for altogether approximately 80 inserts of chloroplast DNA populating the *Haplomitrium ensiformis* mitogenome (see supplementary table 3). Maximum likelihood trees were conducted with IQ-TREE ⁹⁰ after automatic model selection of TIM+F+I+G4 or GTR+F+I+G4 and trees were rooted with the Lindsaeaceae family or the Eupolypod II clade, respectively, for cp1039 and cp2126. Bootstrap support is derived from 500 replicates. **A.** The intergenic region between *nad5* and *sdh4* contains the largest collection of likely independently acquired cpDNA inserts including the largest individual insert cp4165 with 93% similarity to the native chloroplast *ndhH-ndhE* region. The other inserts share variable sequence identities with the native *H. ensiformis* cpDNA ranging from 73% for cp686 to 99% for cp1271. Inserts cp364 and cp749 lack evident homologies in the *H. ensiformis* plastome, but are identified by sequence similarities to cpDNAs in other fern genera like *Asplenium* or *Vittaria*, highlighted in red. **B.** Chloroplast DNA insert cp1039 derived from the chloroplast *atpA-atpF* region is an example for a likely very recently acquired insert as evident from its well-supported sister placement to the newly assembled *H. ensiformis* cpDNA (Fig. 1) in a phylogeny including the homologous plastome regions from diverse polypod ferns. **C.** Chloroplast insert cp2128 embedded in other cp inserts of variable sequence conservation carries a unique morffo element, identified by sequence similarity only in the cpDNA of *Hemionitis subcordata*.

Figure 9. [Rickettsial-like bacterial inserts in the *Haplopteris ensiformis* mitogenome.](#)

Numerous inserts of bacterial, mostly rickettsial, protein-coding sequences were found to be integrated into the *Haplopteris ensiformis* mitogenome. Xenologous bacterial inserts were annotated to indicate their sizes in base pairs, preceded by 'x' (see supplementary table 3), here showing examples for x625 (A), x888 and x1623 (B) and x1920 (c). Average read coverages are shown on top and PCR amplicons used to verify the mitogenome assemblies are indicated, with PCR results exemplarily shown for x625. **A.** Bacterial insert x625 is located next to cpDNA insert cp416 between *cox2* and the downstream part of *rrnL*. Two overlapping PCRs confirm the mitogenome assembly with linkages into both genes for two independent samples from distantly grown *Haplopteris ensiformis* plant isolations but failed to find products for a *Vittaria lineata* sample growing near *H. ensiformis* isolate 1. PCR products of expected sizes (framed with stippled boxes) were cut out and sequenced and confirmed sequence identity with the mitogenome assembly. The graph on top shows a continuity for the average read coverages of ca. 170 x for *cox2*, cp416 and x625 with an increase to ca. 440 x for the downstream *rrnL* region. Insert x625 carries the central region for DNA recombination protein RmuC with the upstream part located on insert x1623. **B.** Xenologous inserts x888 and x1623 are located between *nad1* and cpDNA insert cp232. Coding regions for HscA, an FeS-protein assembly chaperone and for RmuC borne on x1623 (see supplementary table 3 and supplementary figure S3) are truncated, but highly conserved in primary sequence as exemplarily shown in the phylogenetic analysis for RmuC documenting a close association with *Caedimonas varicaedens*. Severely degenerated coding sequences for OMBB, an outer membrane beta-barrel domain containing protein borne on x888 are highly degenerated and do not allow a clear affiliation with a specific Rickettsiales bacterium. **C.** Xenologous insert x1920 represents a continuous stretch of Rickettsia DNA with top similarities to four coding sequences in *Caedimonas* spp. Coding sequences (see supplementary figure S3) are full-length for ribosome maturation factor RimM and the tRNA guanine-methyltransferase TrmD but amino-terminally truncated for MfdD, a transcription-repair-coupling factor and carboxy-terminally truncated for RpL19 encoding protein 19 of the large ribosomal subunit, respectively.

[Supplementary Data](#)

Supplementary Table 1. [Chloroplast RNA editing in *Haplopteris ensiformis*.](#)

The table lists RNA editing sites in chloroplast transcripts of *Haplopteris ensiformis*. The first column indicates positions in the newly assembled cpDNA (accession OM867544), the third column indicates labels for editing sites following a previous nomenclature proposal (Lenz et al., 2010) and the fourth column indicates edits other than those changing codon identities. The second column lists the distance to the respective following editing site with immediately neighboring sites highlighted by

green shading and those only two nucleotides apart with blue shading. Editing site labels are now suggested to be amended for multiple edits affecting single codons (highlighted with yellow shading, see supplementary figure 1). Altogether 443 sites of RNA editing were detected in the *Haplopteris ensiformis* chloroplast transcriptome applying strict criteria for detection. The table also includes cpDNA positions for which RNA editing would have been strongly expected from predictions but for which we could not identify a base conversion. A classification of edits according to genes and quality of codon changes is given under B and C, respectively.

Supplementary Table 2. [Mitochondrial RNA editing in *Haplopteris ensiformis*.](#)

Altogether 1618 sites of RNA editing (1091 sites of C-to-U and 527 of U-to-C editing) were detected in the *Haplopteris ensiformis* mitochondrial transcriptome. Makeup of listing is as in supplementary table 1. Owing to a co-existing pseudogene copy, RNA editing in *atp8* was analyzed by RT-PCR (indicated with “cDNA” in column efficiency).

Supplementary Table 3. [Annotated features in the *Haplopteris ensiformis* mitogenome chromosomes.](#)

List of annotated features (genes, repeat regions, recombination points, promiscuous chloroplast DNA insertions ‘cp’ and bacterial sequence insertions ‘x’) for mtDNA contigs A through L variably integrated into mitogenome chromosomes 1 through 9 as displayed in figures 3 and 4, respectively.

Supplementary Figure S1. Nomenclature extension for multiple edits affecting single codons.

The pipe symbol (|) is added for editing site labels when multiple non-silent C-to-U or U-to-C edits affecting single codons. The respective codon change considering the individual edit alone is given, as usual, at the end. The ultimate codon change outcome when taking also the neighboring non-silent change is additionally indicated before the pipe symbol. **A.** All possible ways of converting YYN (Pro, Leu, Ser and Phe) codons are shown with individual changes in the first or second codon position in the top or bottom lines, respectively. C-to-U editing is shown in blue and U-to-C editing in red. First position edits of CUR or UUR Leucine codons are silent when considered individually but are factually not when accompanied by editing in second codon position (LS|LL and LP|LL, blue shading). We name these events “primary silents”. *Vice versa*, apparent non-silent 1st position edits of CCR proline or UCR serine codons (PL|PS and SL|SP, green shading) may ultimately appear silent when the 2nd position is edited, too. We name these events “secondary silents”. For silent edits in codons also affected by non-silent edits, the codon identities before and after the non-silent edit(s) are shown before and after the underline. **B.** The example shows the exceptional case of five RNA editing events in a row causing Leu-to-Phe exchanges in two successive *nad1* codons. The respective RNA editing efficiencies are indicated. Silent sites are mostly edited with low efficiencies, as here exemplarily seen for nad1eU702LL_SS

edited to only 15%. Notable exceptions are found for efficiently edited silent sites downstream of a Thr-to-Ile codon conversion in *nad4* (C) or for three closely spaced silent edits in the amino-terminal part of the *nad5* coding sequence (D).

Supplementary Figure S2. Prediction of mitochondrial RNA editing.

RNA editing sites were predicted using the “commons” function of PREPACT 3.0⁸⁸ with the selection of reference editomes shown on top (the alga *Chara vulgaris*, the liverwort *Marchantia paleacea*, the moss *Physcomitrella patens*, the lycophytes *Isoetes engelmannii* and *Selaginella moellendorffii*, the eusporangiate ferns *Ophioglossum californicum* and *Psilotum nudum* and the angiosperms *Cocos nucifera* and *Liriodendron tulipifera*). Black font indicates prediction from genomically encoded conserved codons, red font indicates known editing events in the respective reference and single letters indicate a deviating amino acid in a given reference editome. Examples are shown for the *atp9* gene (A) and the *atp6* gene (B).

Supplementary figure S3. [Bacterial protein sequence similarities](#) in the *Haplopteris ensiformis* mitogenome.

Bacterial protein sequence similarities identified in the *Haplopteris ensiformis* mitogenome. Shown are the top similar alignments indicating length of the native protein entry, the numbers of identical and similar residues and total alignment length separated by slashes followed by percentages of identical and similar residues, respectively.

Supplementary figure S4. [Bacterial nucleotide sequence similarities](#) in the *Haplopteris ensiformis* mitogenome.

Alignments exemplarily displaying bacterial nucleotide sequence similarities in the *Haplopteris ensiformis* mitogenome.

References

1. Gualberto, J. M. & Newton, K. J. Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. *Annu Rev Plant Biol* **68**, 225–252 (2017).
2. Arrieta-Montiel, M. P. & Mackenzie, S. A. Plant mitochondrial genomes and recombination. in (ed. Kempken, F.) 65–82 (Springer, 2011).
3. Cole, L. W., Guo, W., Mower, J. P. & Palmer, J. D. High and Variable Rates of Repeat-Mediated Mitochondrial Genome Rearrangement in a Genus of Plants. *Mol Biol Evol* **35**, 2773–2785 (2018).
4. Chevigny, N., Schatz-Daas, D., Lotfi, F. & Gualberto, J. M. DNA Repair and the Stability of the Plant Mitochondrial Genome. *Int J Mol Sci* **21**, 328 (2020).

5. Christensen, A. C. Plant Mitochondria are a Riddle Wrapped in a Mystery Inside an Enigma. *Journal of Molecular Evolution* vol. 89 151–156 Preprint at <https://doi.org/10.1007/s00239-020-09980-y> (2021).
6. Liu, Y. *et al.* The Mitochondrial Genome of the Lycophyte *Huperzia squarrosa*: The Most Archaic Form in Vascular Plants. *PLoS One* **7**, e35168 (2012).
7. Grewe, F., Viehoveer, P., Weisshaar, B. & Knoop, V. A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte *Isoetes engelmannii*. *Nucleic Acids Res* **37**, 5093–5104 (2009).
8. Hecht, J., Grewe, F. & Knoop, V. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biol Evol* **3**, 344–358 (2011).
9. Chaw, S.-M. S. M. *et al.* The Mitochondrial Genome of the Gymnosperm *Cycas taitungensis* Contains a Novel Family of Short Interspersed Elements, Bpu Sequences, and Abundant RNA Editing Sites. *Mol Biol Evol* **25**, 603–615 (2008).
10. Guo, W. *et al.* *Ginkgo* and *Welwitschia* mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. *Mol Biol Evol* **33**, 1448–1460 (2016).
11. Kan, S.-L., Shen, T.-T., Gong, P., Ran, J.-H. & Wang, X.-Q. The complete mitochondrial genome of *Taxus cuspidata* (Taxaceae): eight protein-coding genes have transferred to the nuclear genome. *BMC Evol Biol* **20**, 10 (2020).
12. Sullivan, A. R. *et al.* The Mitogenome of Norway Spruce and a Reappraisal of Mitochondrial Recombination in Plants. *Genome Biol Evol* **12**, 3586–3598 (2020).
13. Jackman, S. D. *et al.* Organellar Genomes of White Spruce (*Picea glauca*): Assembly and Annotation. *Genome Biol Evol* **8**, 29–41 (2015).
14. Putintseva, Y. A. *et al.* Siberian larch (*Larix sibirica* Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome. *BMC Genomics* **21**, 654 (2020).
15. Richardson, A. O., Rice, D. W., Young, G. J., Alverson, A. J. & Palmer, J. D. The ‘fossilized’ mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol* **11**, 29 (2013).
16. Alverson, A. J. *et al.* Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol Biol Evol* **27**, 1436–1448 (2010).
17. Rodríguez-Moreno, L. *et al.* Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. *BMC Genomics* **12**, 424 (2011).
18. Sloan, D. B. *et al.* Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. *PLoS Biol* **10**, e1001241 (2012).
19. Rice, D. W. *et al.* Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* **342**, 1468–1473 (2013).

20. Guo, W., Zhu, A., Fan, W. & Mower, J. P. Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. *New Phytologist* **213**, 391–403 (2017).
21. Knie, N., Grewe, F. & Knoop, V. Monilophyte mitochondrial rps1 genes carry a unique group II intron that likely originated from an ancient paralog in rpl2. *RNA* **22**, 1338–1348 (2016).
22. Zumkeller, S. M., Knoop, V. & Knie, N. Convergent Evolution of Fern-Specific Mitochondrial Group II Intron atp1i361g2 and Its Ancient Source Parologue rps3i249g2 and Independent Losses of Intron and RNA Editing among Pteridaceae. *Genome Biol Evol* **8**, 2505–19 (2016).
23. Knie, N., Grewe, F., Fischer, S. & Knoop, V. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns – a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. *BMC Evol Biol* **16**, 134 (2016).
24. Guo, W., Grewe, F. & Mower, J. P. Variable frequency of plastid RNA editing among ferns and repeated loss of uridine-to-cytidine editing from vascular plants. *PLoS One* **10**, e0117075 (2015).
25. Robison, T. A. *et al.* Mobile Elements Shape Plastome Evolution in Ferns. *Genome Biol Evol* **10**, 2558–2571 (2018).
26. Lenz, H. *et al.* Introducing the plant RNA editing prediction and analysis computer tool PREPACT and an update on RNA editing site nomenclature. *Curr Genet* **56**, 189–201 (2010).
27. Song, M. *et al.* A novel chloroplast gene reported for flagellate plants. *Am J Bot* **105**, 117–121 (2018).
28. Gerke, P. *et al.* Towards a plant model for enigmatic U-to-C RNA editing: the organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort *Anthoceros agrestis*. *New Phytologist* **225**, 1974–1992 (2020).
29. Knie, N., Polsakiewicz, M. & Knoop, V. Horizontal Gene Transfer of Chlamydial-Like tRNA Genes into Early Vascular Plant Mitochondria. *Mol Biol Evol* **32**, 629–34 (2014).
30. Oldenkott, B., Yamaguchi, K., Tsuji-Tsukinoki, S., Knie, N. & Knoop, V. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. *RNA* **20**, 1499–1506 (2014).
31. Bégu, D., Castandet, B. & Araya, A. RNA editing restores critical domains of a group I intron in fern mitochondria. *Curr Genet* **57**, 317–325 (2011).
32. Bégu, D. & Araya, A. The horsetail *Equisetum arvense* mitochondria share two group I introns with the liverwort *Marchantia*, acquired a novel group II intron but lost intron-encoded ORFs. *Curr Genet* **55**, 69–79 (2009).
33. Turmel, M., Otis, C. & Lemieux, C. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *Plant Cell* **15**, 1888–1903 (2003).
34. Suzuki, H. *et al.* Draft genome sequence of *Caedibacter varicaedens*, a kappa killer endosymbiont bacterium of the ciliate *Paramecium biaurelia*. *Genome Announc* **3**, 23–25 (2015).

35. Schuettpeitz, E., Schneider, H., Huiet, L., Windham, M. D. & Pryer, K. M. A molecular phylogeny of the fern family Pteridaceae: assessing overall relationships and the affinities of previously unsampled genera. *Mol Phylogenet Evol* **44**, 1172–1185 (2007).
36. Schuettpeitz, E. & Pryer, K. M. Fern phylogeny inferred from 400 leptosporangiate species and three plastid genes. *Taxon* **56**, 1037–1050 (2007).
37. Schuettpeitz, E. & Pryer, K. M. Fern phylogeny. in *Biology and Evolution of Ferns and Lycophytes* 395–416 (2008).
38. Grusz, A. L., Rothfels, C. J. & Schuettpeitz, E. Transcriptome sequencing reveals genome-wide variation in molecular evolutionary rate among ferns. *BMC Genomics* **17**, 692 (2016).
39. Rothfels, C. J. & Schuettpeitz, E. Accelerated rate of molecular evolution for vittarioid ferns is strong and not driven by selection. *Syst Biol* **63**, 31–54 (2014).
40. Fauskee, B. D., Sigel, E. M., Pryer, K. M. & Grusz, A. L. Variation in frequency of plastid RNA editing within *Adiantum* implies rapid evolution in fern plastomes. *Am J Bot* **108**, 820–827 (2021).
41. Roulet, M. E. *et al.* Multichromosomal structure and foreign tracts in the *Ombrophytum subterraneum* (Balanophoraceae) mitochondrial genome. *Plant Mol Biol* **103**, 623–638 (2020).
42. Kozik, A. *et al.* The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet* **15**, 1–30 (2019).
43. Sloan, D. B. One ring to rule them all? Genome sequencing provides new insights into the ‘master circle’ model of plant mitochondrial DNA structure. *New Phytologist* **200**, 978–985 (2013).
44. Knoop, V. Plant mitochondrial genome peculiarities evolving in the earliest vascular plant lineages. *J Syst Evol* **51**, 1–12 (2013).
45. Mower, J. P., Sloan, D. B. & Alverson, A. J. Plant Mitochondrial Genome Diversity: The Genomics Revolution. in *Plant Genome Diversity Volume 1* vol. 1 123–144 (Springer Vienna, 2012).
46. Sessa, E. B. *et al.* Between two fern genomes. *Gigascience* **3**, 15 (2014).
47. Li, F.-W. *et al.* Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature* **4**, 460–472 (2018).
48. Huang, X. *et al.* The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat Plants* **8**, 500–512 (2022).
49. Fang, Y. *et al.* The genome of homosporous maidenhair fern sheds light on the euphyllphyte evolution and defences. *Nature Plants* 2022 1–14 (2022) doi:10.1038/s41477-022-01222-x.
50. Marchant, D. B. *et al.* Dynamic genome evolution in a model fern. doi:10.1038/s41477-022-01226-7.
51. Wickell, D. *et al.* Underwater CAM photosynthesis elucidated by *Isoetes* genome. *Nat Commun* **12**, 1–13 (2021).
52. Mower, J. P. Variation in protein gene and intron content among land plant mitogenomes. *Mitochondrion* **53**, 203–213 (2020).

53. Guo, W., Zhu, A., Fan, W., Adams, R. P. & Mower, J. P. Extensive shifts from cis to trans splicing of gymnosperm mitochondrial introns. *Mol Biol Evol* **37**, 1615–1620 (2020).
54. Malek, O., Brennicke, A. & Knoop, V. Evolution of trans-splicing plant mitochondrial introns in pre-Permian times. *Proc Natl Acad Sci U S A* **94**, 553–558 (1997).
55. Malek, O. & Knoop, V. Trans-splicing group II introns in plant mitochondria: the complete set of cis-arranged homologs in ferns, fern allies, and a hornwort. *RNA* **4**, 1599–609 (1998).
56. Burger, G., Yan, Y., Javadi, P. & Lang, B. F. Group I intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. *Trends in Genetics* **25**, 381–386 (2009).
57. Pombert, J. F. & Keeling, P. J. The mitochondrial genome of the entomoparasitic green alga *Helicosporidium*. *PLoS.One.* **5**, e8954 (2010).
58. Pelin, A. *et al.* The mitochondrial genome of the arbuscular mycorrhizal fungus *Gigaspora margarita* reveals two unsuspected trans -splicing events of group I introns. *New Phytologist* **194**, 836–845 (2012).
59. Qiu, Y.-L. *et al.* The deepest divergences in land plants inferred from phylogenomic evidence. *Proc Natl Acad Sci U S A* **103**, 15511–15516 (2006).
60. Cho, Y., Qiu, Y. L., Kuhlman, P. & Palmer, J. D. Explosive invasion of plant mitochondria by a group I intron. *Proc Natl Acad Sci U S A* **95**, 14244–14249 (1998).
61. Cusimano, N., Zhang, L. B. L.-B. & Renner, S. S. Reevaluation of the *cox1* group I intron in Araceae and angiosperms indicates a history dominated by loss rather than horizontal transfer. *Mol Biol Evol* **25**, 265–276 (2008).
62. Stern, D. B. & Lonsdale, D. M. Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. *Nature* **299**, 698–702 (1982).
63. Wickell, D. A. & Li, F. On the evolutionary significance of horizontal gene transfers in plants. *New Phytologist* **225**, 113–117 (2020).
64. Gao, J.-G. Tracking the evolutionary innovations of plant terrestrialization. *Gene* 145203 (2020) doi:10.1016/j.gene.2020.145203.
65. Mower, J. P., Stefanović, S., Young, G. J. & Palmer, J. D. Plant genetics: gene transfer from parasitic to host plants. *Nature* **432**, 165–166 (2004).
66. Davis, C. C., Anderson, W. R. & Wurdack, K. J. Gene transfer from a parasitic flowering plant to a fern. *Proceedings of the Royal Society B: Biological Sciences* **272**, 2237–2242 (2005).
67. Bergthorsson, U., Adams, K. L., Thomason, B. & Palmer, J. D. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424**, 197–201 (2003).
68. Won, H. & Renner, S. S. Horizontal gene transfer from flowering plants to *Gnetum*. *Proc Natl Acad Sci U S A* **100**, 10824–10829 (2003).
69. Mower, J. P., Jain, K. & Hepburn, N. J. The Role of Horizontal Transfer in Shaping the Plant Mitochondrial Genome. *Adv Bot Res* **63**, 41–69 (2012).
70. Taylor, Z. N., Rice, D. W. & Palmer, J. D. The Complete Moss Mitochondrial Genome in the Angiosperm *Amborella* Is a Chimera Derived from Two Moss Whole-Genome Transfers. *PLoS One* **10**, e0137532 (2015).

71. Sinn, B. T. & Barrett, C. F. Ancient Mitochondrial Gene Transfer between Fungi and the Orchids. *Mol Biol Evol* **37**, 44–57 (2020).
72. Schrallhammer, M., Castelli, M. & Petroni, G. Phylogenetic relationships among endosymbiotic R-body producer: Bacteria providing their host the killer trait. *Syst Appl Microbiol* **41**, 213–220 (2018).
73. Grosser, K. *et al.* More than the ‘killer Trait’: Infection with the Bacterial Endosymbiont *Caedibacter taeniospiralis* Causes Transcriptomic Modulation in *Paramecium* Host. *Genome Biol Evol* **10**, 646–656 (2018).
74. Davis, M., Ying, Z., Brunner, B., Pantoja, A. & Ferwerda, F. Rickettsial relative associated with papaya bunchy top disease. *Curr Microbiol* **36**, 80–84 (1998).
75. Yurchenko, T. *et al.* A gene transfer event suggests a long-term partnership between eustigmatophyte algae and a novel lineage of endosymbiotic bacteria. *ISME J* **12**, 2163–2175 (2018).
76. Yang, A., Narechania, A. & Kim, E. Rickettsial endosymbiont in the “early-diverging” streptophyte green alga *Mesostigma viride*. *J Phycol* **52**, 219–229 (2016).
77. Siguier, P., Gournayre, E. & Chandler, M. Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiol Rev* **38**, 865–891 (2014).
78. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
79. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
80. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res* **9**, gkw955 (2016).
81. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–52 (2011).
82. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512 (2013).
83. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
84. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* **44**, (2016).
85. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. in *Statistical Genomics. Methods in Molecular Biology* (eds. Mathé, E. & Davis, S.) vol. 1418 283–334 (Humana Press, 2016).
86. Piechotta, M., Wyler, E., Ohler, U., Landthaler, M. & Dieterich, C. JACUSA: site-specific identification of RNA editing events from replicate sequencing data. *BMC Bioinformatics* **18**, 7 (2017).

87. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res* **47**, W59–W64 (2019).
88. Lenz, H., Hein, A. & Knoop, V. Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in PREPACT 3.0. *BMC Bioinformatics* **19**, 255 (2018).
89. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–5 (2009).
90. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).

Figure 1
A

A



A. *Haplopteris ensiformis* reveals a typical plant circular plastome structure consisting of a large (LSC) and a small (SSC) single copy region separated by a pair of inverted repeats (IR) and an expectedly

conserved, ancestral gene and intron complement. The genome map was created using OGDRAW 87. Gene categories are indicated in the legend. Numbers in parentheses indicate the amount of C-to-U (blue) and U-to-C (red) RNA editing for the respective genes. Creations of start or stop codons by C-to-U editing are indicated by symbols '>' and '*' and the removal of stop codons by U-to-C editing is indicated by the exclamation marks, respectively. B. The cpDNA of *Haplopteris elongata* (accession MH173086) features two "morffo" elements ("mobile ORFs in fern organelles") in the IR region between *rrn5* and the 3'-part of the trans-splicing *rps12* gene. Recognizable sequence homologues of morffo2 can presently only be identified in *Cyclosorus interruptus* (accession MN599066, Thelypteridaceae) and *Histiopteris incisa* (accession MH319942, Dennstaedtiaceae) and a homologue of morffo1 (orange) can presently only be found in the distant fern *Hymenophyllum holochilum* (accession MH265124, Hymenophyllales). Only the upstream part of morffo2 (378 bp) is present in the *H. ensiformis* plastome, while a cpDNA insert in its mitogenome contains an extended region of 628 bp.

Figure 2



Figure 2

The accD gene example for chloroplast RNA editing in *Haplopteris ensiformis*.

A. Sequence alignment of the *H. ensiformis* accD gene below its homologue in *Psilotum nudum* shown as one selected example reference out of 22 used for prediction of RNA editing. Alignment was created by PREPACT 88 with identical nucleotides and amino acid shown in grey font and predicted C-to-U RNA editing in blue and reverse U-to-C editing in red. Codons framed by boxes were confirmed as editing sites. The stippled rectangle highlights potential reverse edit accDeC580FL remaining unconfirmed but strongly suggested by the chloroplast editome references in PREPACT 3.0 with the exception of *Ginkgo biloba*.

coding and tRNA gene sequence are given in lighter and darker blue, respectively, and introns are indicated with black arrows. Numbers next to genes indicate C-to-U (blue) and U-to-C RNA edits (red) with additional symbols indicating removal (!) or creation of stop (*) or start (>) codons, respectively. Introns are indicated with stippled lines and additional black arrows and their standardized labels. Ribosomal RNA genes (here on contig A) are shown in red with the *rrnL* gene featuring a complex gene structure requiring trans-splicing via the disrupted group I intron *rrnLi825g1* and PSX labels indicate pseudogene fragments shown in grey. The peculiar case of the trans-spliced group I intron *rrnLi825g1* is highlighted in yellow. Genes for tRNAs of chloroplast or bacterial origin are indicated in green or purple, respectively. For clarity, no other chloroplast or bacterial DNA insertions are shown here. The latter are listed together with annotated features in supplementary table 3.

D



The *Haplopteris ensiformis* mtDNA contigs D-L are linked to recombination points in chromosome 5 (Fig. 3) or within themselves, creating further and/or alternative mtDNA arrangements. Recombination

endpoints are labeled and the display of contigs and labels for genes and RNA editing events is as in figure 3. Possible, circular chromosomal structures chr1 to chr4 and chr6 to chr9 are shown.

Figure 5

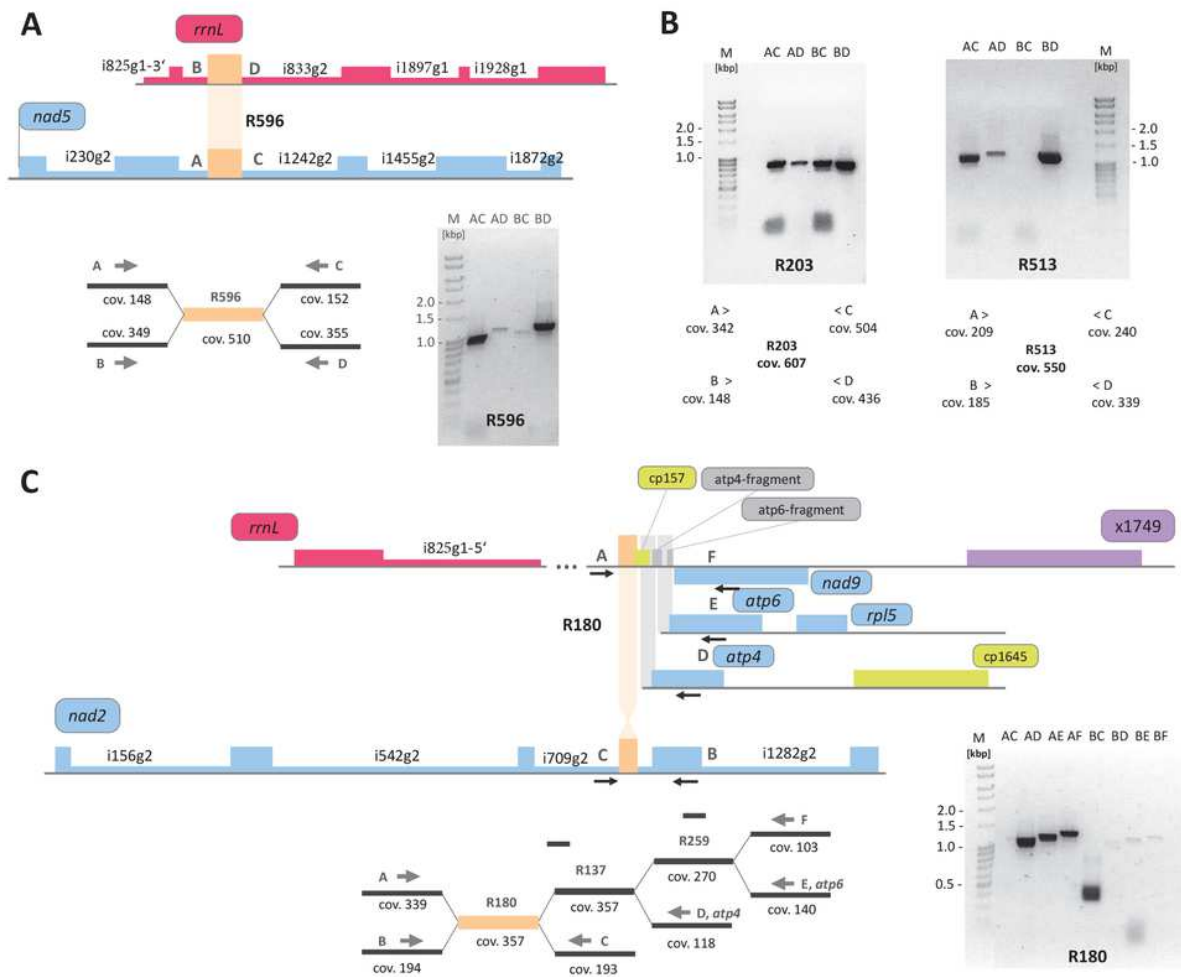


Figure 5

Repeats and recombination in the *Haplopteris ensiformis* mitogenome.

Recombinations across repeats (in orange) R596 (A), repeats R203 and R 513 (B) and R180 (C) was investigated by ts-PCR ("template-switch-avoiding") strategies. A. R596 is identically present in domains

Mitochondrial RNA editing in *Haplopteris ensiformis*: the rpl6-rps13-rps11 case.

The alignment exemplarily shows RNA editing heavily affecting the rpl6-rps13-rps11 co-transcript including the removal of seven stop codons within the first 20 codons of the rps11 reading frame. An internal PCR amplicon covers 46 editing sites from rpl6eC106*R to rps11eC136*Q. Synthesis of cDNA was primed either with random hexamers (n6) or with specific primers covering the end of rps11 including the stop codon generation and two edits in the 3'-UTR in an edited (P+) or unedited (P-) version. The alignment displays edits (C-to-U in blue and U-to-C in red) clearly revealed in the RT-PCRs primed by the three different approaches and the results from the RNA-Seq data (RRM). Silent codon edits shown below the protein alignment were exclusively identified in the latter.

Figure 7

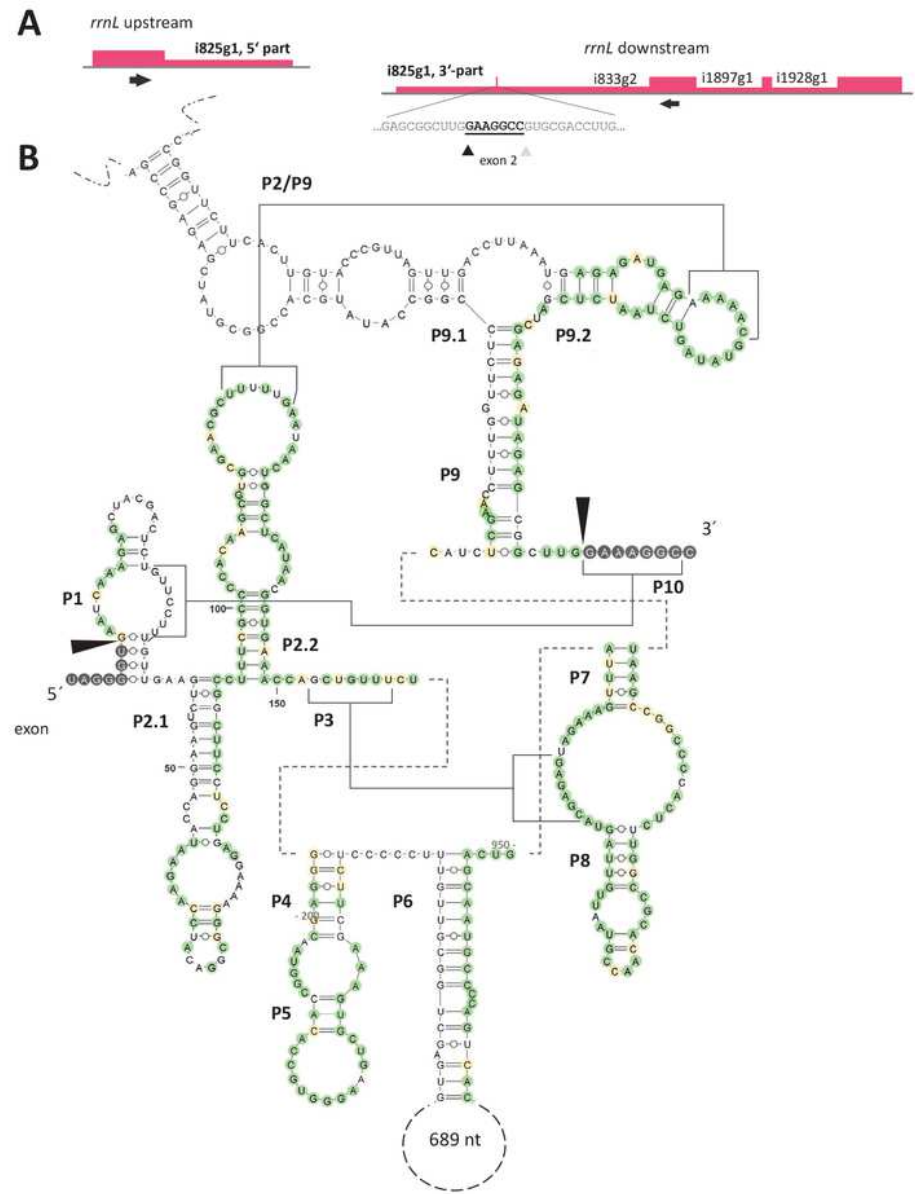


Figure 7

A trans-splicing group I intron in the *Haplopteris ensiformis* mitochondrial *rrnL* gene.

A. Maturation of *rrnL* includes trans-splicing of the disrupted group I intron *rrnLi825g1*. Three additional introns are removed from the downstream part of *rrnL*. Group II intron *rrnLi833g2* has a homologue conserved in the mtDNAs of liverworts and the downstream group I introns *rrnLi897g1* and *rrnLi1928g1*

are conserved in other Polypodiales species. Splice sites of *rrnLi825g1* and *rrnLii833g2* (black and grey triangles, respectively) frame the tiny second *rrnL* exon of only eight nucleotides. B. Secondary structure model of the disrupted, trans-splicing group I intron *rrnLi825g1*. The figure was generated using the VARNA software 89. Regions for base-pairing between the upstream and downstream parts of *rrnLi825g1* are found in the group I intron secondary structure P9.1. Intron *rrnLi825g1* intron has a positional orthologue in the sweet-water alga *Chara vulgaris*. Green and yellow shading indicates identical nucleotides and transitions, respectively.

Figure 8

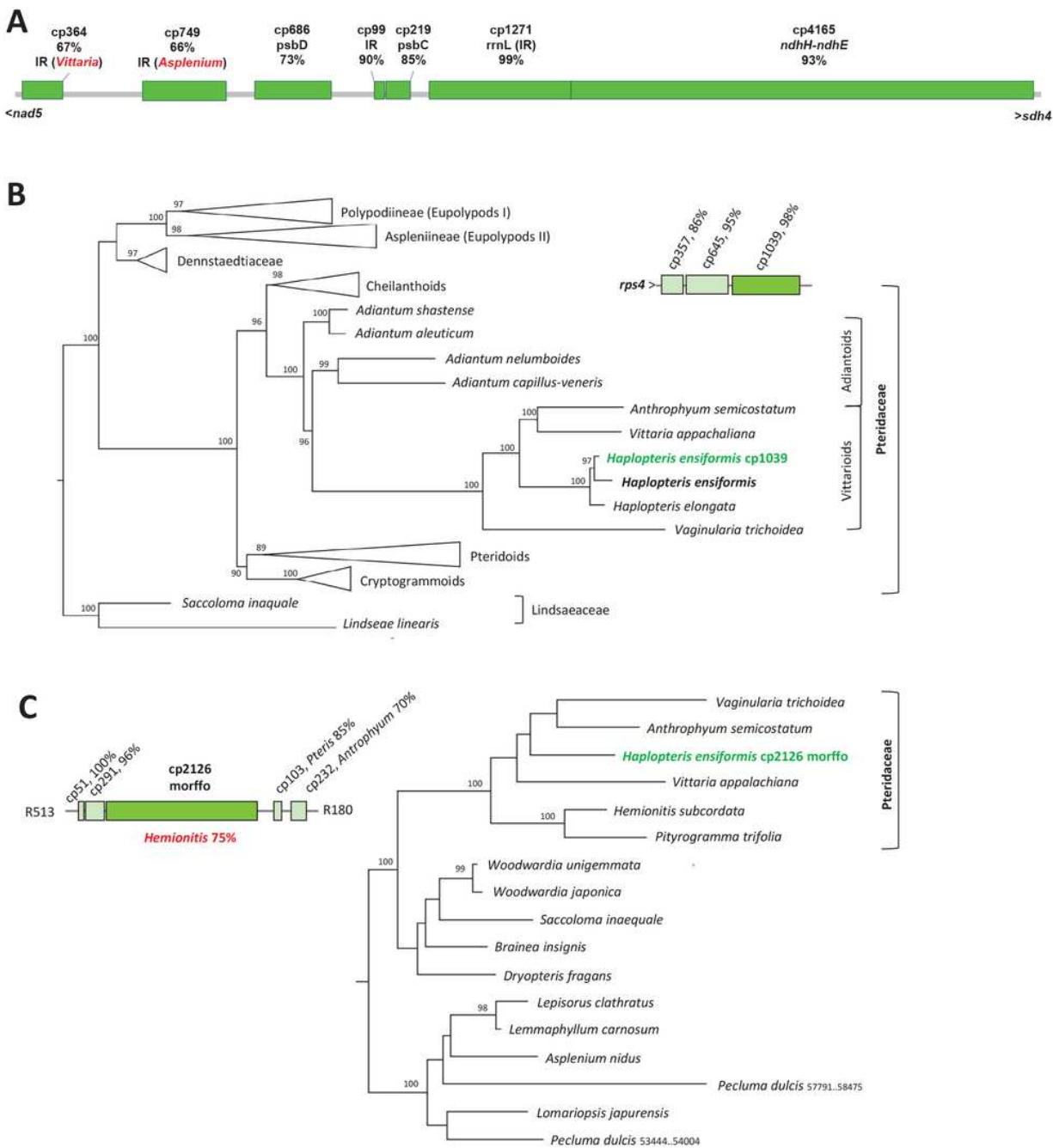


Figure 8

A multitude of chloroplast DNA inserts in the *Haplopteris ensiformis* mitogenome.

Selected examples for altogether approximately 80 inserts of chloroplast DNA populating the *Haplopteris ensiformis* mitogenome (see supplementary table 3). Maximum likelihood trees were conducted with IQ-TREE 90 after automatic model selection of TIM+F+I+G4 or GTR+F+I+G4 and trees were rooted with the Lindsaeaceae family or the Eupolypod II clade, respectively, for cp1039 and cp2126. Bootstrap support is derived from 500 replicates. A. The intergenic region between *nad5* and *sdh4* contains the largest collection of likely independently acquired cpDNA inserts including the largest individual insert cp4165 with 93% similarity to the native chloroplast *ndhH-ndhE* region. The other inserts share variable sequence identities with the native *H. ensiformis* cpDNA ranging from 73% for cp686 to 99% for cp1271. Inserts cp364 and cp749 lack evident homologies in the *H. ensiformis* plastome, but are identified by sequence similarities to cpDNAs in other fern genera like *Asplenium* or *Vittaria*, highlighted in red. B. Chloroplast DNA insert cp1039 derived from the chloroplast *atpA-atpF* region is an example for a likely very recently acquired insert as evident from its well-supported sister placement to the newly assembled *H. ensiformis* cpDNA (Fig. 1) in a phylogeny including the homologous plastome regions from diverse polypod ferns. C. Chloroplast insert cp2128 embedded in other cp inserts of variable sequence conservation carries a unique morffo element, identified by sequence similarity only in the cpDNA of *Hemionitis subcordata*.

Figure 9

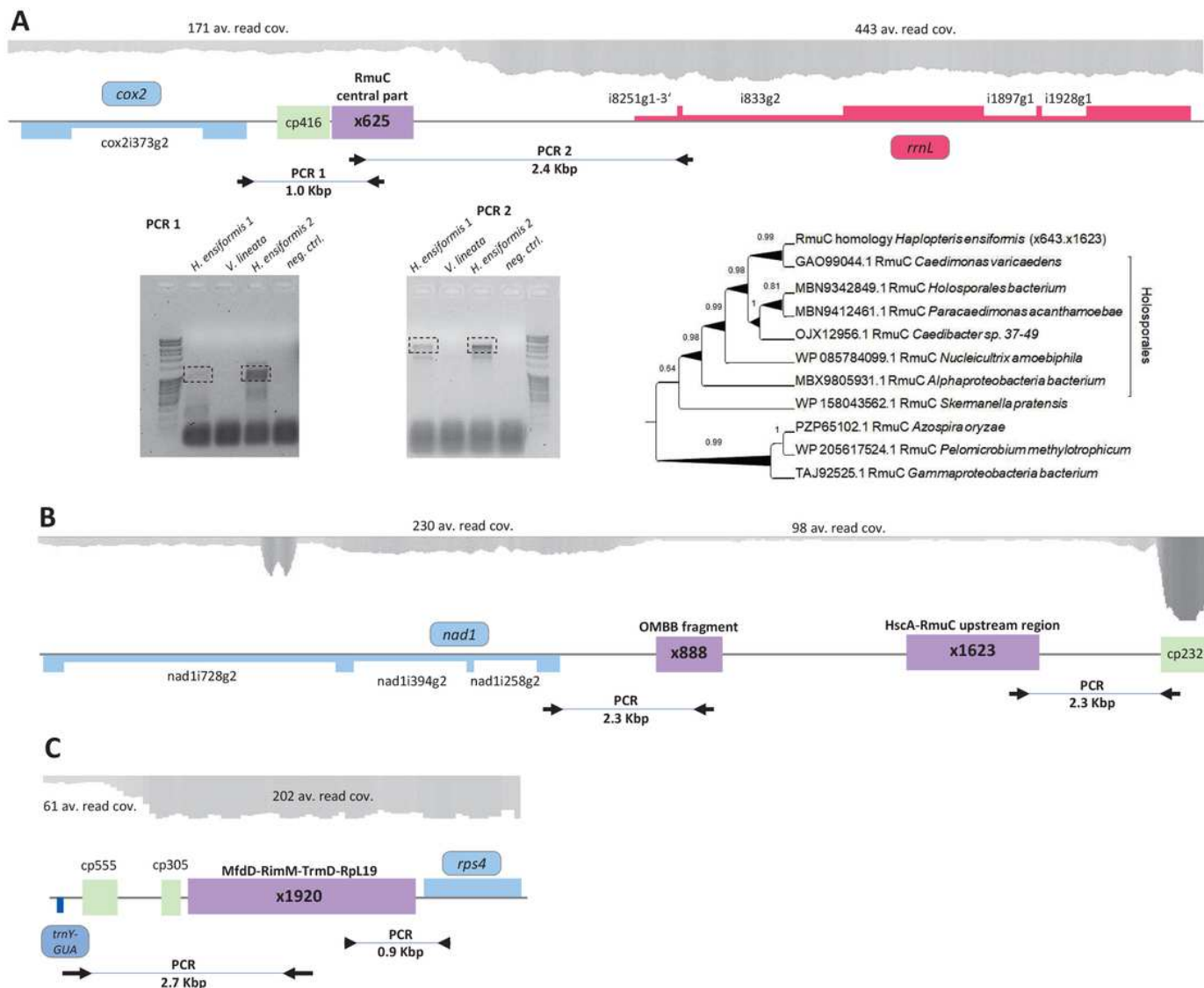


Figure 9

Rickettsial-like bacterial inserts in the Haplopteris ensiformis mitogenome.

Numerous inserts of bacterial, mostly rickettsial, protein-coding sequences were found to be integrated into the Haplopteris ensiformis mitogenome. Xenologous bacterial inserts were annotated to indicate their sizes in base pairs, preceded by 'x' (see supplementary table 3), here showing examples for x625 (A), x888 and x1623 (B) and x1920 (c). Average read coverages are shown on top and PCR amplicons used to verify the mitogenome assemblies are indicated, with PCR results exemplarily shown for x625. A. Bacterial insert x625 is located next to cpDNA insert cp416 between cox2 and the downstream part of rrnL. Two overlapping PCRs confirm the mitogenome assembly with linkages into both genes for two independent samples from distantly grown Haplopteris ensiformis plant isolations but failed to find products for a Vittaria lineata sample growing near H. ensiformis isolate 1. PCR products of expected

sizes (framed with stippled boxes) were cut out and sequenced and confirmed sequence identity with the mitogenome assembly. The graph on top shows a continuity for the average read coverages of ca. 170 x for *cox2*, *cp416* and *x625* with an increase to ca. 440 x for the downstream *rrnL* region. Insert *x625* carries the central region for DNA recombination protein RmuC with the upstream part located on insert *x1623*. B. Xenologous inserts *x888* and *x1623* are located between *nad1* and *cpDNA* insert *cp232*. Coding regions for HscA, an FeS-protein assembly chaperone and for RmuC borne on *x1623* (see supplementary table 3 and supplementary figure S3) are truncated, but highly conserved in primary sequence as exemplarily shown in the phylogenetic analysis for RmuC documenting a close association with *Caedimonas varicaedens*. Severely degenerated coding sequences for OMBB, an outer membrane beta-barrel domain containing protein borne on *x888* are highly degenerated and do not allow a clear affiliation with a specific Rickettsiales bacterium. C. Xenologous insert *x1920* represents a continuous stretch of *Rickettsia* DNA with top similarities to four coding sequences in *Caedimonas* spp. Coding sequences (see supplementary figure S3) are full-length for ribosome maturation factor RimM and the tRNA guanine-methyltransferase TrmD but amino-terminally truncated for MfdD, a transcription-repair-coupling factor and carboxy-terminally truncated for Rpl19 encoding protein 19 of the large ribosomal subunit, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplFigure1ZumkelleretalHaplopterismtDNANC20220919final2.pdf](#)
- [SupplFigure2ZumkelleretalHaplopterismtDNANC20220919final2.pdf](#)
- [SupplFigure3ZumkelleretalHaplopterismtDNANC20220919final2.pdf](#)
- [SupplFigure4ZumkelleretalHaplopterismtDNANC20220919final2.pdf](#)
- [SupplTab1ZumkelleretalCPEditingfinalp1.pdf](#)
- [ZumkelleretalSupplTab2MTEditingcomplete.pdf](#)
- [SupplTab3Zumkelleretal2022insertsfeaturesSZ20220727.pdf](#)