This version of the article has been accepted for publication, after peer review and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s10265-018-1077-y

TITLE PAGE:

Full plastome sequence of the fern *Vandenboschia speciosa* (Hymenophyllales): structural singularities and evolutionary insights

Ruiz-Ruano F.J.; Navarro-Domínguez, B.; Camacho J.P.M.; Garrido-Ramos M.A.
Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain

Abstract

We provide here the first full chloroplast genome sequence, i.e., the plastome, for a species belonging to the fern order Hymenophyllales. The phylogenetic position of this order within leptosporangiate ferns, together with the general scarcity of information about fern plastomes, places this research as a valuable study on the analysis of the diversity of plastomes throughout fern evolution. Gene content of *V. speciosa* plastome was similar to that in most ferns, although there were some characteristic gene losses and lineage-specific differences. In addition, an important number of genes required U to C RNA editing for proper protein translation and two genes showed start codons alternative to the canonical AUG (AUA). Concerning gene order, *V. speciosa* shared the specific 30-kb inversion of euphyllophytes plastomes and the 3.3-kb inversion of fern plastomes, keeping the ancestral gene order shared by eusporangiate and early leptosporangiate ferns. Conversely, *V. speciosa* has expanded IR regions comprising the *rps7*, *rps12*, *ndhB* and *trnL* genes in addition to rRNA and other tRNA genes, a condition shared with several eusporangiate ferns, lycophytes and hornworts, as well as most seed plants.

Keywords: ferns, gene losses, IR expansion, non-canonical start codons, plastome, RNA editing, *Vandenboschia speciosa*

Introduction

The era of genomics allows not only the untangling of the most recondite details of nuclear genome composition in eukaryotic cells, but also facilitates the study of the genomes of cytoplasmic organelles in a relatively simple and time-saving way. Eukaryotic cells contain mitochondria responsible for cellular respiration and cellular metabolism regulation. Specifically, plants are characterized by also having chloroplasts that allow plants to convert light energy into chemical energy through photosynthesis. Both organelles are prokaryotic endosymbionts in origin and both contain their own genomes, which have been drastically reduced during evolution. Current organelle genomes are basically composed of genes related with their function in the eukaryotic cell, so that the plastome contains genes that encode for proteins involved in photosynthesis and for proteins and RNAs involved in gene expression.

Plastomes of land plants are generally conserved in structure, gene content and gene order (Wicke et al. 2011; Wolf et al. 2010; Green 2011; Gao et al. 2010; Wolf and Karol 2012; Ruhlman and Jansen 2014; Xu et al. 2015). Most plastomes are characterized by a quadripartite structure, including two copies of an inverted repeat (IRA and IRB) and the large (LSC), and small (SSC) single copy regions. The plastome usually includes 120-130 genes and varies in size from 120 to 170 Kilobases (kb). However, specific lineages of land plants are characterized by gene and/or intron losses or duplications changing gene content in the plastome as well as some rearrangements changing its gene order (Wicke et al. 2011; Wolf et al. 2010; Green 2011; Gao et al. 2010; Wolf and Karol 2012; Ruhlman and Jansen 2014; Xu et al. 2015). These changes appear to be phylogenetically informative (Wolf et al. 2010; Gao et al. 2009, 2011, 2013; Karol et al. 2010; Grewe et al. 2013; Kim et al. 2014; Zhu et al. 2016; Logacheva et al. 2017) and some examples are known in bryophytes (e.g. Wolf and Karol 2012; Park et al. 2018), lycophytes (e.g. Wolf et al. 2005; Tsuji et al. 2007; Guo et al. 2016), monilophytes (e.g., Wolf et al. 2003, 2010; Roper et al. 2007; Gao et al. 2009, 2011, 2013; Karol et al. 2010; Grewe et al. 2013; Kim et al. 2014; Zhong et al. 2014; Lu et al. 2015; Zhu et al. 2016; Logacheva et al. 2017), and seed plants (in both gymnosperm and angiosperm lineages) (e.g. Wicke et al. 2011; Green 2011; Ruhlman and Jansen 2014; Xu et al. 2015; Zhu et al. 2016; Sun et al. 2016).

Plastome sequence is thus an useful tool to build phylogenies which have contributed very much our understanding of plant evolution. It is thus essential to analyze plastome diversity across all plant lineages, and this is especially crucial in understudied groups such as ferns (Roper et al. 2007; Wolf et al. 2010; Karol et al. 2010; Gao et al. 2013; Grewe et al. 2013; Kim et al. 2014; Labiak and Karol 2017). Extant ferns are a lineage of non-seed vascular plants including 45 families and about 280 genera (PPG I 2016; Schuettpelz and Pryer 2007; Rai and

Graham 2010; Lehtonen 2011; Rothfels et al. 2015; Knie et al. 2015; Qi et al. 2018). Molecular phylogenies have revealed a basal dichotomy within the vascular plants (Pryer et al. 2001), separating the lycophytes (less than 1% of the current vascular plants) from the euphilophytes, among which there are two major clades, the monilophytes or ferns and the spermatophytes or seed plants. Monilophyte includes the classes Psilotopsida (orders Psilotales and Ophioglossales), Equisetopsida (order Equisetales), Marattiopsida (order Marattiales) and Polypodiopsida (Figure 1). The class Polypodiopsida, also known as leptosporangiate ferns (sporangia arise from a single epidermal cell and not from a group of cells as in eusporangiate ferns), includes the vast majority of extant ferns, with more than 80% of about 10,500 fern species. Conversely, the classes Psilotopsida, Equisetopsida and Marattiopsida comprise the grade of eusporangiate ferns. The monophyletic group of leptosporangiates is composed of a total of seven orders (PPG I 2016; Schuettpelz and Pryer 2007; Rai and Graham 2010; Lehtonen 2011; Rothfels et al. 2015; Knie et al. 2015; Qi et al. 2018). Three leptosporangiate orders, Salviniales (heterosporous ferns), Cyatheales (tree ferns), and Polypodiales (polypods), form the large monophyletic clade of core leptosporangiates (Smith et al. 2006). The rest of leptosporangiate species are included within the orders Osmundales, Hymenophyllales, Gleicheniales and Schizaeales (Figure 1).

Among the nearly 2,000 complete plastome sequences available in the GenBank database, only a few tens belong to monilophytes (Kim et al. 2014; Lu et al. 2015; Zhu et al. 2016; Logacheva et al. 2017; our own look-over to the NCBI web page https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/). In addition, the information about fern plastomes is limited to a few representatives of most, but not all, fern orders. And, within the different orders, only a few representative species have been analyzed from some families (Lu et al. 2015). This clearly indicates that the diversity of plastome structures in ferns has insufficiently been explored, precluding a comprehensive study of plastome evolution in this group (Grewe et al. 2013; Logacheva et al. 2017). Furthermore, among complete plastome sequences, there are no representatives from the order Hymenophyllales, a key lineage within non-core, early leptosporangiates. Thus, the recent phyloplastomic analysis of ferns carried out by Lu et al. (2015) included at least one representative species of each Monilophyte order except for the order Hymenophyllales.

Our aim here is to fill this gap by analyzing the full sequence of the plastome of *Vandenboschia speciosa* (Hymenophyllaceae, Hymenophyllales). The family Hymenophyllaceae, with about 600 species, has its origin in the Triassic and its highest degree of diversification took place about 160 mya during the Jurassic (Pryer et al. 2004). This family is characterized by the "filmy" aspect of its sporophytes (fronds constituted by a translucent

sheet with a single layer of cells). Although this family shows pantropical distribution, with high diversity in morphology and habitat occupation (Dubuisson et al. 2003; Ebihara et al. 2007), *V. speciosa* constitutes a rare Macaronesian-European endemism, as it is the only representative in the area and one of the most vulnerable fern species in Europe, restricted to places considered as refuges of tertiary flora. The two phases of its life cycle are perennial and capable of reproducing by vegetative propagation (Rumsey et al. 1999). The sporophyte is rhizomatous and able to spread by fragmentation of its rhizome. The gametophyte is very different from the typical heart-shaped prothallus, it is characterized by being epigeous and narrowly talose or filamentous (to such an extent that it is often confused with the protonema of a bryophyte), and produces geminiferous buds. All these characteristics add valuable scientific interest to the selection of this species as representative of the order Hymenophyllales for plastome sequencing.

Materials and Methods

Materials

Vandenboschia speciosa specimens were collected at the Canuto de Ojén-Quesada (OJEN) population located in the Alcornocales Natural Park (Cádiz, Spain), geographical coordinates: N36.127°/W5.585° (for voucher reference, see Ben-Menni Schuler et al., 2017). Sporophytes were frozen in liquid nitrogen in the field and were stored at −80°C. Genomic DNA (gDNA) was isolated from five individuals of this population using the DNeasy plant Mini kit (Quiagen). A pool of DNA was generated from this five specimen DNAs and Next Generation Sequencing of the pool was carried out based on the Illumina HiSeq 2000 PE 2x101 nt yielding about 16 Gb data. 16 Gb represent a ~1,5x coverage of the nuclear genome (10.496 Gb; Obermayer et al. 2002) and ~923x coverage of the chloroplast genome (146.874 bp and 0,85% of the Illumina reads; see also Figure S1). Illumina sequencing data can be accessed at Short Read Archive (SRA) Genbank database in the BioProject PRJNA387541 under the accession number SRX2844191.

Genome assembly

The absence of the complete sequence of a plastome in the GenBank for a relative species, except two partial sequences belonging to *Vandenboschia radicans* of about 16 kb (Wolf et al. 2010; accession number HM021800) and 9 kb (Gao et al. 2011; accession number HQ658104), lead us to get a reference sequence of *V. speciosa* to be used as a seed in later steps. For this, we performed a clustering of Illumina reads and a *de novo* assembly of repetitive elements by means of the RepeatExplorer (RE) pipeline (Novák et al. 2013). This

software groups similar reads in clusters, assembles the reads of each cluster to generate contigs for repetitive elements, and annotates the contigs. We ran RE with 2x500,000 *V. speciosa* gDNA Illumina reads and then selected the longest contigs showing the highest number of reads annotated as "plastid". We then checked the homology of this sequence with other plant plastomes performing a BLASTN (Altschul et al. 1990) search against the NCBI's NR database. Once confirmed the homology with plastomes, we used this partial sequence as a seed to assembly the remaining sequence of the *V. speciosa* plastome using a random selection of 2x2,000,000 reads with the MITObim v1.8 software (Hahn et al. 2013) with the "--quick" option. A single run of MITObim was unable to assembly the full sequence, for which reason we used the assembled sequence for a new run with this same software in order to lengthen the assembly. We ran MITObim four times to complete the full sequence.

We assessed the quality of the resulting assembly by mapping the Illumina reads on it with the Bowtie2 (Langmead and Salzberg2012) software considering only read pairs completely mapped (--no-mixed) and maximum insert size of 1000 nt (--maxins 1000). Then, we plotted the coverage along the whole plastome using the output of Pysamstats (https://github.com/alimanfoo/pysamstats) and manually checked the uniformity of the mappings and the presence of complete reads in truncated genes using IGV (Thorvaldsdóttir et al. 2013), a visualization tool for genome-scale datasets (Figure S1).

Annotation and related studies

Annotation of the *V. speciosa* plastome was performed using a combination of GeSeq-Annotation of Organellar Genomes (Tillich et al. 2017) and DOGMA (Dual Organellar GenoMe Annotator) (Wyman et al. 2004). From this initial annotation, putative starts, stops, and intron positions were determined by comparisons with homologous genes in other plastomes and by considering the possibility of RNA editing, which can modify the start and stop positions. tRNA genes were annotated using GeSeq, DOGMA, tRNAscan-SE v2.0 (Lowe and Chan 2016) and ARAGORN (Laslett et al. 2004). The circular gene map of the *V. speciosa* plastome was drawn by OGDRAW (Lohse et al. 2013) located at the GeSeq plataform (https://chlorobox.mpimp-golm.mpg.de/index.html) and subsequent minor manual modification to indicate the quadripartite structure of the plastome.

Plastome structure, gene content, and other general characteristics were then compared with the published fern plastomes available in the NCBI website (http://www.ncbi.nlm.nih. gov/) and in the literature.

The EMBOSS suite of bioinformatics tools (Rice et al. 2000) was used for the detection of short internal repeats (direct or inverted), as well as palindromes, in the intergenic spacer

sequences, using the programs MATCHER, ETANDEM, EINVERTED, POLYDOT and PALINDROME.

The annotated plastome of *V. speciosa* was deposited in GenBank under accession number MH648610.

3. Results and Discussion

3.1. Assembly and Genome structure

The RE analysis clusterized 49% of the Illumina reads in 538 clusters, eight of which were annotated as chloroplast DNA sequences. We selected the cluster no. 217, composed of 815 reads and its longest contig with 4626 nt to be used as a seed for MITObim assembly. After four consecutive MITObim runs, we obtained the full plastome sequence. Bowtie2 mapping yielded a uniform profile across all the sequence.

The plastome of *V. speciosa* consists of 146,874 bp (Figure S2) and shows a typical quadripartite structure including a large single-copy (LSC) region of 89,620 bp and a small single-copy (SSC) region of 21,398 bp, separated by a pair of identical inverted repeats (IRA and IRB) of 17,928 bp each (Figure 2). The overall structure of the *V. speciosa* plastome is typical of vascular plants although several rearrangements were detected (see below). Table 1 compares the structure of *V. speciosa* plastome with that of different fern species, as well as those of lycophytes and bryophytes. Sequenced fern plastomes have lengths ranging between 131.760 bp (*Equisetum hyemale*) and 157.260 bp (*Lygodium japonicum*), likewise in other land plant groups (Table 1). Specifically, within this range, the *V. speciosa* plastome shows an intermediate size.

Plastome size is a feature that depends on lineage-specific expansions and contractions (see Table 1). More than half of the *V. speciosa* plastome is composed of non-coding regions (79,835 bp; 55.36%) including introns and intergenic sequences, and has an overall G+C content of 37.5%, similar to other vascular plant plastomes, but higher than that in non-vascular land plant plastomes (Table 1) (reviewed in Park et al. 2018). As in other cases (for example, see Gao et al. 2009), the G+C content is unevenly distributed across the plastome depending on location and functional group (Figure 2): 54.76% in rRNA genes, 53.62% in tRNA genes, 38.40% in protein coding regions and 34.46% in intergenic spacers.

3.2. Gene content

Table 2 lists the genes found in the plastome of *V. speciosa*, a repertoire of genes typical of land plant plastomes. We identified four rRNAs genes (*rrn4.5*, *rrn5*, *rrn16* and *rrn23*, all of them being duplicated within the IR regions), 31 tRNAs genes (six of them being duplicated within the IR regions) and 85 protein-coding genes, three of which are duplicated within the IR regions and four are truncated at the 5' end suggesting a process of pseudogenization (Table 2). As mentioned above, Bowtie2 mapping yielded a uniform profile across all the sequence and we did not find truncated reads in regions comprising these four genes, meaning that their truncations were not assembling artifacts (Figure S1). There are 17 intron-containing genes, including six tRNA genes and 11 protein-coding genes, summing up 21 introns since three genes have more than one intron (Table 2). *rps12* is a trans-spliced gene showing two exons, one located in the LSC region (3' end) and the other placed, in two copies, within the IRs (5' end).

From an evolutionary point of view, *V. speciosa* plastome shows a gene content that is characteristic of early leptosporangiate ferns (Kim et al 2014). Notwithstanding, there were a few singularities in this species concerning some tRNA and protein-coding genes (Figure 3) examined in detail in the following subsections.

tRNA genes

The set of 31 tRNAs genes of the *V. speciosa* plastome is assumed to be sufficient for the translation of chloroplast mRNAs (Gao et al. 2009). However, it lacks the tRNA gene for lysine (*trnK*). Thus, although chloroplast tRNAs could read most codons by using two-out-of-three and wobble mechanisms (Pfitzinger et al. 1990; Gao et al. 2009), the absence of *trnK* genes for reading any of the two lysine codons suggests that cytosolic tRNAs may be imported into chloroplasts, despite a lack of experimental evidence (Gao et al. 2009). The *trnK*-UUU gene of land plants has a large intron in which is nested the *matK* gene, a gene that encodes the protein maturase involved in splicing type II introns (Kuo et al. 2011; Wicke and Quandt 2009). This structure is conserved in eusporangiate ferns as well as in the leptosporangiate *Osmunda cinnamomea* plastome (Osmundaceae, Osmundales). However, the *trnK*-UUU gene was lost in *V. speciosa* whereas the *matK* gene was still maintained, likewise in the remaining leptosporangiate ferns. This supports the loss of *trnK*-UUU after the divergence of Osmundales (Kuo et al. 2011; Grewe et al. 2013) (Figure 3), but not the loss of the intron-encoded *matK*. In plants, *matK* preferentially catalyzes splicing of the trnK intron but, in ferns, it may also have maintained an ancient and generalized function as a maturase that catalyzes the splicing reactions of other group II introns in the chloroplast genome, even after the loss of its co-evolved group II intron splicing (Hausner et al. 2006; Duffy et al. 2009).

In addition, *V. speciosa* plastome lacks two more tRNA genes (*trnS*-CGA and *trnT*-UGU) which are usually present in the plastomes of other land plants (Figure 3). The loss of *trnS*-CGA is shared with other leptosporangiates except, again, with *Osmunda cinnamomea* which still preserves this gene (Kim et al. 2014; Grewe et al. 2013; Gao et al. 2013). The *trnT*-UGU gene was also lost independently in the eusporangiate *Ophioglossum californicum* (Grewe et al. 2013; Gao et al. 2013), but not in other eusporangiates such as *Equisetum*, *Psilotum* or *Angiopteris* (Grewe et al. 2013; Karol et al. 2010; Roper et al. 2007).

Other specific feature is that the *trnL*-CAA gene, placed close to the *ndhB* gene, has mutated in *V. speciosa* to *trnL*-CAG while, in most leptosporangiate ferns, it has been lost, although it is still conserved in Gleicheniales (Kim et al. 2014; Grewe et al. 2013). Conversely, the *trnL*-UAA gene, placed between the *rps4* and the *ndhJ* genes, has mutated to *trnL*-CAA after the divergence of Osmundales (Kim et al. 2014), including *V. speciosa* (this paper).

Figure 3 shows that the loss of the *trnV*-GAC gene occurred in the common ancestor of core leptosporangiates and the Schizaeales (Grewe et al. 2013; Kim et al. 2014; Labiak and Karol 2017) but not in the rest of leptosporangiates represented by *V. speciosa*. On the other hand, the *trnR*-CCG gene is preserved in all leptosporangiate ferns but has undergone several sequential anticodon changes, in some species, leading to alternative anticodons, pseudogenes or even *trnR*-UCA genes (Wolf et al. 2003, 2004; Grewe et al. 2013). It has been suggested that *trnR*-UCA might recognize internal UGA stop codons acting as a failsafe mechanism to ensure that arginine is correctly inserted into the protein at any internal UGA codons that were not properly converted by U-to-C RNA editing into CGA (Grewe et al. 2013). However, that is not the case for *V. speciosa* which has UGA stop codons interrupting 12 genes, but lacks the *trnR*-UCA mutation (see section 3.3).

Protein-coding genes

Protein-coding gene content of *V. speciosa* plastome is similar to that of most ferns, which conserve about 86 of these genes, although there are some lineage-specific differences (Figure 3). In the case of *V. speciosa*, we have found the partial loss of the sequences of four protein-coding genes (*ycf1*, *ycf2*, *rps16* and *rpl22*). We assume that these genes are going through a process of pseudogenization that probably would lead to their complete lost. In fact, we have found a few additional indels (short deletions of one or a few nucleotides) both in *ycf1* and *ycf2* genes although not in the other two cases. These four genes have important roles in chloroplast gene expression and function; therefore, it is likely that these functions are supplied by other means such as those found in other species with missing plastid genes.

Thus, the *rps16* has been lost in several lineages of ferns such as Ophioglossales, Psilotales and Equisetales (eusporangiate ferns) (Grewe et al. 2013; Kim et al. 2014) and it seems that it will be lost in one specific lineage of leptosporangiates (*V. speciosa*) (Figure 3). Furthermore, several parallel losses of the *rps16* gene have occurred along land plant phylogeny (Xu et al. 2015). Missing genes in specific plant lineages might have been relocated from the plastome to the nuclear genome or, alternatively, these functions were supplied by other means (Gantt et al. 1991; Gao et al. 2009; Jansen et al. 2011; Ruhlman and Jansen 2014; Keller et al. 2017; Park et al. 2018). In the genus *Lupinus* (Leguminosae), the functional role of this gene in the chloroplast has been replaced by the nuclear *rps16* gene (Keller et al. 2017; Ruhlman and Jansen 2014). And it has been shown that the *rpl22* gene has been transferred to the nucleus in different flowering plant lineages (Gantt et al. 1991; Jansen et al. 2011; Ruhlman and Jansen 2014).

The pseudogenization of the *ycf1* gene has also been documented in *Angiopteris evecta* (Roper et al. 2007). Both *ycf1* and *ycf2* genes have been designated as essential genes because their targeted disruption results in unstable mutant phenotypes (Drescher et al. 2000; Ruhlman and Jansen 2014). However, both *ycf1* and *ycf2* genes appear to be missing or pseudogenized in distinct land plant lineages (Ruhlman and Jansen 2014). The *ycf66* gene have been lost in many fern species (Gao et al. 2011), including *V. speciosa* (this paper) although it is present in the partial sequence of the plastome of *V. radicans* (Gao et al. 2011). In fact, we have detected a residual fragment of this gene in the plastome of *V. speciosa*. In addition, this fragment has stops codons and several indels. Specifically, this gene has been lost or is pseudogenized in at least four times in fern lineages Ophioglossales, Psilotales, Equisetales, and core leptosporangiates, suggesting that the occurrence of this gene is highly unstable and, probably, that this ORF is irrelevant for an hypothetical and, up to now, uncharacterized, protein.

Other lineage-specific protein-coding gene losses included in Figure 3 were not found in *V. speciosa*: the *psaM* gene is missing from all sequenced plastomes of the order Polypodiales (Wolf et al. 2003; Gao et al. 2013; Grewe et al. 2013), and it was lost in parallel also from some species in other lineages (Labiak and Karol 2017). Events of recurrent losses were also found for other genes such as the *chl* genes (*chlB*, *chlL* and *chlN*), which were lost in *Psilotum nodum* (Psilotales) (Grewe et al. 2013) and *Actinostachys pennula* (Schizaeaceae, Schizaeales) (Labiak and Karol 2017), but also in some angiosperm plastomes (Chaw et al. 2004; Jansen et al. 2007).

The *ndhB* gene merits here a specific commentary as it may be duplicated or not depending on whether it is located within the IR or the LSC regions. Furthermore, according to

Gao et al. (2013), there are three types of *ndhB* genes in ferns: one copy of exon 1 plus one copy of exon 2 (1:1 type), one copy of exon 1 plus two copies of exon 2 (1:2 type), and two copies of exon 1 plus two copies of exon 2 (2:2 type) (Gao et al. 2013). In heterosporous ferns (core leptosporangiates) *ndhB* is single copied and located in the LSC (1:1 type) (see Gao et al. 2013). However, the second exon is duplicated in the plastome IRs of the two other orders of core leptosporangiates, the tree ferns and the polypod ferns (1:2 type; see Gao et al. 2013). *ndhB* gene resides in IRs in two out of the four orders comprising the eusporangiate ferns, the psilotales and the marattioids (2:2 type) (Gao et al. 2013). In contrast, the *ndhB* is 1:1 type in ophioglossoids and equisetales (Gao et al. 2013). Remarkably, the copy status of *ndhB* in the non-core leptosporangiate fern lineages is highly variable, as *V. speciosa* (Hymenophyllales) shows a 2:2 type *ndhB* gene (this study), while *Osmunda cinnamomea* (Osmundales) shows a 1:1 type, but *Diplopterygium glaucum* (Gleicheniales) and *Lygodium japonicum* (Schizaeales) a 1:2 type (Gao et al. 2013). These data support the existence of parallel changes among major fern lineages that have occurred in a region (IR) prone to rearrangements and expansions/contractions (see next section).

Recently, Song et al. (2018) have reported a novel chloroplast protein-coding gene (*ycf94*) of probable functional significance in ferns and, potentially, bryophytes and lycophytes. This ORF is located between the *rps16* and the *matK* genes. We have not found evidence for the presence of this ORF in the intergenic sequence between *rps16* and *matK* genes in *V. speciosa*.

3.3. Non-canonical start codons and internal stop codons

Examination of start and stop codons (internal and terminal) in the *V. speciosa* plastome resulted in 27 protein-coding genes with one (most of the 27 genes) or a few internal stop codons that required U-to-C RNA editing to overcome the premature end translation by the incorporation of the appropriate amino acid (Table 3). U-to-C editing in the first codon position has been proved that might be involved in the repair of internal stop codons to either arginine or glutamine (Kugita et al. 2003; Wolf et al. 2004; Sugiura 2008; Guo et al. 2015; Knie et al. 2016). In fact, in *V. speciosa*, the stop codons came from mutations from CGA (arginine) to UGA or from CAA (glutamine) to UAA, according to the results of BLAST search among homologous genes from other fern species.

U-to-C editing appears to be restricted in occurrence in most land plants but, it has been found that, in some fern lineages, U-to-C editing even exceeds the canonical C-to-U editing (Knie et al. 2016). In contrast, the reverse process, RNA editing by C-to-U conversions to reconstitute conserved codon identities has been extensively documented in ferns (Wolf et

al. 2004) and is common in land plant chloroplasts and mitochondria, except in some liverworts which have secondarily lost this type of RNA editing (Sugiura 2008; Guo et al. 2015; Knie et al. 2016; Groth-Malonek et al. 2007; Rüdinger et al. 2012).

In addition to the safeguard role against premature stopping, RNA editing occasionally can create an AUG initiation codon from an ACG codon by C-to-U edit (Wolf et al. 2004; Sugiura 2008). We have found that the *atpH*, the *ndhG* and *ndhH* the genes of *V. speciosa* begin with ACG. But we also found two other genes beginning with non-canonical AUG start codon (AUA), specifically the *petD* and the *rps15* genes (Table 3). It is known that translation can initiate at codons other than AUG (Kearse and Wilusz 2017) and that some organelle and prokaryote genomes use GUG and AUA as alternate start codons for some of these genes (Wolf et al. 2004; Kuroda et al. 2007; Guo et al. 2015).

3.4. Gene order

There are four key rearrangements that characterize the evolution of fern plastomes and mark differences between fern and the rest of land plants as well as among different fern groups (Gao et al. 2011).

First, euphyllophytes (ferns and seed plants) have differentiated plastomes with respect to lycophytes and non vascular plants (hornworts, mosses and liverworts) (Raubeson and Jansen 1992). The rearrangement consists in a 30-kb inversion comprising the region between the *ycf2* and the *psbM* genes (Figure 4). As expected, *V. speciosa* shares this syntenic rearranged region with the rest of ferns.

Second, monilophytes (ferns) have a unique rearranged region consisting in a 3.3-kb inversion known as the *trnG/trnT* inversion in which the genes between these two tRNA genes have the *trnG-psbZ-trnS-psbC-psbD-trnT* gene order while in the rest of land plants, the order is the inverse (Wolf et al. 2003, 2010). Also expectedly, *V. speciosa* has the *trnG-psbZ-trnS-psbC-psbD-trnT* gene order characteristic of ferns (Figure 4).

Third, according to gene order, the fern plastomes can be classified into two main types (Gao et al. 2009, 2011, 2013; Zhu et al. 2016) (Figures 4 and 5). One comprising the plastomes of taxa diversifying before the separation of the Schizaeales, which shares the ancestral gene order (Wolf et al. 2010), and the other comprising the plastomes of Schizaeales and of core leptosporangiates, which has a derived gene order (Wolf et al. 2010; Gao et al. 2013; Labiak and Karol 2017). The derived gene order consists of highly rearranged inverted repeats (IRs) with the rRNA genes arranged in reverse order in comparison to all other plants, generated by two partially overlapping inversions spanning LSC and IR regions (Hasebe et al. 1990, 1992; Wolf et al. 2003, 2010) (Figure 5). An additional difference between the ancestral

and derived gene order is the region comprised between the *rpoB* and *psbZ*, which includes the *trnD* inversion (Figure 4). This characteristic gene order is exclusive of Polypodiopsida (Gao et al. 2009).

Inverted Repeat (IRs) Region Expansions

The presence of inverted regions (IRs) within the plastomes is a common feature to most plants with a few exceptions among seed plants (Guisinger et al. 2011; Guo et al. 2014; Zhu et al. 2016). This region comprises mostly four genes for the four chloroplastidial ribosomal RNAs, several tRNA genes and, depending of the species, a few protein-coding genes. Figure 5 shows an inference of the ancestral inverted repeat (IR) content during land plant evolution, based on a previous proposal by Zhu et al. (2016), but with some modifications. Our table includes not only the *V. speciosa* data but also the reordered IR region of Polypodiopsida and that of Schizaeales. The rearrangement consisted in a pair of partially overlapping inversions in the region of the inverted repeat that occurred in the common ancestor of Schizaeales and Polypodiopsida (most ferns) (Wolf et al. 2010). Figure 5 includes the structure of these ferns in a separate frame because these lineage-specific changes comprising the inversion of the IR gene order have no accommodation in the ancestral reconstruction. Furthermore, the IRs of Polypodiales plastomes are dynamic, driven by such events as gene loss, duplication and putative lateral transfer from mitochondria (Logacheva et al. 2017).

As can be seen in Figure 5, seed plants (both, gymnosperms and angiosperms) show a derived condition apparently resulting from an expansion of the IRs that include not only rRNA and tRNA genes but also protein coding genes commonly found in the LSC region in ferns, lycophytes and bryophytes. Furthermore, there are some specific gains and/or losses in certain lineages. Major losses occurred in *Erodium* and *Medicago* or in Pinaceae or Cupressophytes where plastomes retains only residual inverted repeats or lack IRs (reviewed in Zhu et al. 2015). Some lycophytes and fern lineages have also expanded IRs like seed plants, although in a less extension. For example, this occurred in *V. speciosa* (this paper), but also in some eusporangiate ferns as well as in some lycophytes and non vascular plants such as hornworts. The length of the expansion depends of the species. In *V. speciosa* for example, the expansion includes the *rps7*, *rps12* and *ndhB* genes, which is the more common situation in some eusporangiate ferns and in lycophytes as well as hornworts with expanded IRs. Interestingly, this gene order and organization at IRs, define the plastome of *V. speciosa* as one of the species that experienced the recurrent IR expansion. Notwithstanding, the possibility remains that what we have defined as parallel expansions of the IR regions might be the result of a

preservation of the ancestral condition while recurrent contractions might explain the smallest inverted regions found in non seed plants. This is also supported by the fact that the IRs of Polypodiales, even reordered, also contains the protein-coding genes found in expanded IRs (Figure 5).

3.5. Repeat sequences in the plastome of V. speciosa

Table 4 shows all the tandem and inverted short repeats found within the intergenic spacers of the *V. speciosa* plastome. There are 2 direct repeats and 5 inverted repeats. The two direct repeats are interestingly localized in the *rpoB/psbZ* region. Specifically, one of them (a 27-bp repeat) was previously described in *V. radicans* (Gao et al 2011). Likewise *V. speciosa* (this study; see Figure S2), *V. radicans* shows a 27-bp tandem repeat within the *trnY-trnE* intergenic spacer which shows strong sequence similarity with the anticodon domain of *trnY*, the tRNA gene for tyrosine (Gao et al 2011). While the 27-bp repeat of *V. radicans* is repeated 17 times expanding approximately 3-fold in length the spacer (Gao et al 2011), this repeat appears only 5 times in the *V. speciosa* plastome (this paper). These repeats have the potential to form stem-loop structures which point to a significant functional relevance and to play a major role in the expansion of the trnY-trnE intergenic spacer (Gao et al. 2011).

In addition, the plastome of *V. speciosa* contains one more tandem direct repeat, being short (6 bp) and repeated 12 times in the petN-psbM intergenic spacer (also the rpoB/psbZ region). Extensive rearrangements in the plastome of different species have been associated with the concentration of direct repeats and tRNA genes (Chumley et al. 2006; Haberle et al. 2008; Guisinger et al. 2011). In fact, the *rpob/psbZ* region is characterized by notable structural rearrangements through fern evolution directly related with the accumulation of tRNA genes as well as by highly variable intergenic spacers in some lineages which includes gene losses and intergenic expansions throughout direct repeats (Gao et al. 2011).

Additionally, the *V. speciosa* plastome shows five short inverted repeats in five different intergenic spacers (Table 4). Pairwise comparisons between the two components of four of the inverted repeats revealed 100% identity, while the identity was 84% between the two components of the fifth inverted repeat. Additionally, this latter inverted repeat is duplicated because it is within the IR regions between the *ndhB* and the *trnL*-CAG genes. Inverted repeats have the potential to form cruciform structures, which are fundamentally important for a wide range of biological processes, including replication, regulation of gene expression, nucleosome structure and recombination (Brázda et al. 2011) and might be involved in plastome rearrangements (Kolb et al. 2009; Inagaki et al. 2013).

Acknowledgments: This research has been financed by the Spanish Ministerio de Economía y Competitividad and FEDER founds, grant: CGL2010-14856 (subprograma BOS). The Dirección General de Gestión del Medio Natural y Espacios Protegidos of the Consejería de Medio Ambiente y Ordenación del Territorio de la Junta de Andalucía authorized and facilitates the sampling of the material. We are highly indebted to Carmen Rodríguez Hiraldo and to Jaime Pereña Ortiz who, together with the team of Agentes de Medio Ambiente of the Consejería, helped us with the sampling procedure.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Ben-Menni Schuler SBM, García-López MC, López-Flores I, Nieto-Lugilde M, Suárez-Santiago VN (2017) Genetic diversity and population history of the Killarney fern, *Vandenboschia speciosa* (Hymenophyllaceae), at its southern distribution limit in continental Europe. Bot J Linn Soc 183: 94-105.

Brázda et al (2011) Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol Biol 12: 33.

Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58: 424-441 .

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23: 2175-2190.

Dubuisson J-Y, Hennequin S, Douzery EJP, Cranfill RB, Smith AR, Pryer KM (2003) *rbcL* phylogeny of the fern genus *Trichomanes* (Hymenophyllaceae), with special reference to neotropical taxa. Int J Plant Sci 164: 753-761.

Duffy AM, Kelchner SA, Wolf PG (2009) Conservation of selection on *matK* following an ancient loss of its flanking intron. Gene 438: 17-25.

Drescher A, Ruf S, Calsa T, Carrer H, Bock R (2000) The two largest chloroplast genome encoded open reading frames of higher plants are essential genes. Plant J 22: 97-104

Ebihara A, Iwatsuki K, Ito M, Hennequin S, Dubuisson J-Y (2007) A global molecular phylogeny of the fern genus *Trichomanes* (Hymenophyllaceae) with special reference to stem anatomy. Bot J Linnean Society 155: 1-27.

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD (1991) Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J 10: 3073-3078.

Gao L, Su Y-J, Wang T (2010) Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. J Syst Evol 48: 77-93.

Gao L, Wang B, Wang Z-W, Zhou Y, Su Y-J, Wang T (2013) Plastome sequences of *Lygodium japonicum* and *Marsilea crenata* reveal the genome organization transformation from basal ferns to core leptosporangiates. Genome Biol Evol 5: 1403-1407.

Gao L, Yi X, Yang Y-X, Su Y-J, Wang T (2009) Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. BMC Evolutionary Biology 9: 130.

Gao L, Zhou Y, Wang Z-W, Su Y-J, Wang T (2011) Evolution of the rpoB-psbZ region in fern plastid genomes: notable structural rearrangements and highly variable intergenic spacers. BMC Plant Biology 11: 64.

Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. Plant J 66: 34-44.

Grewe F, Guo W, Gubbels EA, Hansen AK, Mower JP (2013) Complete plastid genomes from *Ophioglossum californicum*, *Psilotum nudum*, and *Equisetum hyemale* reveal an ancestral land plant genome structure and resolve the position of Equisetales among monilophytes. BMC Evolutionary Biology 13: 8.

Groth-Malonek M, Wahrmund U, Polsakiewicz M, Knoop V (2007) Evolution of a pseudogene: exclusive survival of a functional mitochondrial nad7 gene supports *Haplomitrium* as the earliest liverwort lineage and proposes a secondary loss of RNA editing in Marchantiidae. Mol Biol Evol 24: 1068-1074.

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28: 583-600.

Guo W, Grewe F, Cobo-Clark A, Fan W, Duan Z, Adams RP, Schwarzbach AE, Mower JP (2014) Predominant and substoichiometric isomers of the plastid genome coexist within *Juniperus* plants and have shifted multiple times during cupressophyte evolution. Genome Biology and Evolution 6: 580-590.

Guo W, Grewe F, Mower JP (2015) Variable frequency of plastid RNA editing among ferns and repeated loss of uridine-to-cytidine editing from vascular plants. PLoS ONE 10: e0117075.

Guo Z-Y, Zhang H-R, Shrestha N, Zhang X-C (2016) Complete chloroplast genome of a valuable medicinal plant, *Huperzia serrata* (Lycopodiaceae), and comparison with its congener. Applications in Plant Sciences 4: 1600071.

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. J Mol Evol 66:350-361.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res 41, e129.

Hasebe M, Iwatsuki K (1990) Chloroplast DNA from *Adiantum capillus-veneris* L., a fern species (Adiantaceae); clone bank, physical map and unusual gene localization in comparison with angiosperm chloroplast DNA. Curr Genet 17: 359-364.

Hasebe M, Iwatsuki K (1992) Gene localization on the chloroplast DNA of the maiden hair fern, *Adiantum capillus-veneris*. J Plant Res 105:413-419.

Hausner G, Olson R, Simon D, Johnson I, Sanders ER, Karol KG, McCourt RM, Zimmerly S (2006) Origin and evolution of the chloroplast *trnK* (*matK*) intron: a model for evolution of group II intron RNA structures. Mol Biol Evol 23: 380-391.

Inagaki H, Ohye T, Kogo H, Tsutsumi M, Kato T, Tong M, Emanuel BS, Kurahashi H (2013) Two sequential cleavage reactions on cruciform DNA structures cause palindrome-mediated chromosomal translocations. Nature Communications 4: 1592.

Jansen RK, Cai Z, Raubeson LA et al (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369-19374.

Jansen RK, Saski C, Lee S, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three rosids (*Castanea* , *Prunus* , *Theobroma*): evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol 28: 835-847.

Karol KG, Arumuganathan K, Boore JL, Duffy AM, Everett KDE, Hall JD, Hansen SK, Kuehl JV, Mandoli DF, Mishler BD, Olmstead RG, Renzaglia KS, Wolf PG (2010) Complete plastome sequences of *Equisetum arvense* and *Isoetes flaccida*: implications for phylogeny and plastid genome evolution of early land plant lineages. BMC Evolutionary Biology 10: 321.

Kearse MG, Wilusz JE (2017) Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes and Development 31: 1717-1731.

Keller J, Rousseau-Gueutin M, Martin G.E., Morice J., Boutte J., Coissac E., Ourari M., Aïnouche M., Salmon A., Cabello-Hurtado F., Aïnouche A (2017) The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from *Lupinus*. DNA Research 24: 343-358.

Kim HT, Chung MG, Kim K-J (2014) Chloroplast genome evolution in early diverged leptosporangiate ferns. Mol Cells 37: 372-382.

Knie N, Grewe F, Fischer S, Knoop V (2016) 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 Evolutionary Biology 16: 134.

Knie N, Fischer S, Grewe F, Polsakiewicz M, Knoop V (2015) Horsetails are the sister group to all other monilophytes and Marattiales are sister to leptosporangiate ferns. Mol Phyl Evol 90: 140-149.

Kolb J, Chuzhanova NA, Högel J, Vasquez KM, Cooper DN, Bacolla A, Kehrer-Sawatzki H (2009) Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. Chromosome Res 17: 469-483.

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res 31: 2417–2423

Kuo LY, Li FW, Chiou WL, Wang CN (2011) First insights into fern matK phylogeny. Mol Phyl Evol 59: 556-566.

Kuroda H, Suzuki H, Kusumegi T, Hirose T, Yukawa Y, Sugiura M (2007) Translation of psbC mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine-Dalgarno sequence in tobacco chloroplasts. Plant Cell Physiol 48: 1374-1378.

Labiak PH, Karol KG (2017) Plastome sequences of an ancient fern lineage reveal remarkable changes in gene content and architecture. Am J Bot 104: 1008-1018.

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357-359.

Laslett D, Canback B (2004) ARAGORN, a program for the detection of transfer RNA and transfer-messenger RNA genes in nucleotide sequences. Nucleic Acids Res 32: 11-16

Lehtonen S (2011) Towards resolving the complete fern tree of life. PLoS ONE 6: e24851.

Logacheva MD, KrinitsinaAA, Belenikin MS, Khafizov K, Konorov EA, Kuptsov SV, Speranskaya AS (2017) Comparative analysis of inverted repeats of polypod fern (Polypodiales) plastomes reveals two hypervariable regions. Plant Biology 17:255.

Lohse M, Drechsel O, Kahlau S and Bock R (2013) OrganellarGenomeDRAW - a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41: W575-W581.

Lowe TM, Chan PP (2016) tRNAscan-SE on-line: search and contextual analysis of transfer RNA genes. Nucleic Acids Res 44: W54-W57.

Lu J-M, Zhang N, Du X-Y, Wen J, Li D-Z (2015) Chloroplast phylogenomics resolves key relationships in ferns. J Syst Evol 53: 448-457.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29: 792-793.

Obermayer R, Leitch IJ, Hanson L, Bennett MD (2002) Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. Ann Bot 90: 209-217.

Park M, Park H, Lee H, Lee B, Lee J (2018) The complete plastome sequence of an Antarctic bryophyte *Sanionia uncinata* (Hedw.) Loeske. Int J Mol Sci 19: 709.

Pfitzinger H, Weil JH, Pillay DT, Guillemaut P (1990) Codon recognition mechanisms in plant chloroplasts. Plant Mol Biol 14: 805-814.

PPG I (2016) A community-derived classification for extant lycophytes and ferns. J Syst Evol 54: 563-603.

Pryer KM, Schuettpelz E, Wolf PG, Schneider H, Smith AR, Cranfill R (2004) Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. Am J Bot 91: 1582-1598.

Pryer KM, Schneider H, Smith AR, Cranfill R, Wolf PG, Hunt JS, Sipes SD (2001) Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature, 409: 618-622.

Qi X et al (2018) A well-resolved fern nuclear phylogeny reveals the evolution history of numerous transcription factor families. Mol Phyl Evol 127: 961-977.

Rai HS, Graham SW (2010) Utility of a large, multigene plastid data set in inferring higher-order relationships in ferns and relatives (monilophytes). Am J Bot 97: 1444-1456.

Raubeson LA, Jansen RK (1992) Chloroplast genomes of plants. In Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Edited by: Henry RJ. London: CABI Publishing; pp: 45-68.

Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276-277.

Roper JM, Hansen SK, Wolf PG, Karol KG, Mandoli DF, Everett KDE, Kuehl J, Boore JL (2007) The Complete Plastid Genome Sequence of *Angiopteris evecta* (G. Forst.) Hoffm. (Marattiaceae). Am Fern J 97: 95-106.

Rothfels KJ et al. (2015) The evolutionary history of ferns inferred from 25 low-copy nuclear genes. Am J Bot 102: 1089-1107.

Rüdinger M, Volkmar U, Lenz H, Groth-Malonek M, Knoop V (2012) Nuclear DYW-type PPR gene families diversify with increasing RNA editing frequencies in liverwort and moss mitochondria. J Mol Evol 74: 37-51.

Ruhlman TA, Jansen RK (2014) Pal Maliga (ed.), Chloroplast Biotechnology: Methods and Protocols, Methods in Molecular Biology, vol. 1132. Springer Science + Business Media New York 2014.

Rumsey FJ, Vogel JC, Russell SJ, Barrett JA, Gibby M (1999) Population genetics and conservation biology of the endangered fern *Trichomanes speciosum* (Hymenophyllaceae) in Scotland. Biol J Linnean Soc 66: 333-344.

Schuettpelz E, Pryer KM (2007) Fern phylogeny inferred from 400 leptosporangiate species and three plastid genes. Taxon 56: 1037-1050.

Smith AR, Pryer KM, Schuettpelz E, Korall P, Schneider H, Wolf PG (2006) A classification for extant ferns. Taxon 55: 705-731.

Song M, Kuo L-Y, Huiet L, Pryer KM, Rothfels CJ, Li F-W (2018. A novel chloroplast gene reported for flagellate plants. Am J Bot 105: 117-121.

Sugiura M (2008) RNA Editing in Chloroplasts. In: H.U. Göringer (ed.), RNA Editing. Nucleic Acids and Molecular Biology 20. Springer-Verlag Berlin Heidelberg 2008.

Sun Y, Moore MJ, Zhang S, Soltis PS, Soltis DE, Zhao T, Meng A, Li X, Li J, Wang H (2016) Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. Mol Phyl Evol 96: 93-101.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14: 178-192.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45: W6-W11.

Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K (2007) The chloroplast genome from a lycophyte (microphyllophyte), *Selaginella uncinata*, has a unique inversion, transpositions and many gene losses. J Plant Res 120: 281-290.

Wicke S, Quandt D (2009) Universal primers for the amplification of the plastid *trnK/matK* region in land plants. Anales Jard Bot Madrid 66: 285-288.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76: 273-297.

Wolf PG, Karol KG (2012) Plastomes of bryophytes, lycophytes and ferns. Pages 89 – 102 In Advances in Photosynthesis and Respiration. Vol 35. “Genomics of chloroplasts and mitochondria”. Editors: Ralph Bock and Volker Knoop. Springer, Dordrecht.

Wolf PG, Karol KG, Mandolib DF, Kuehld J, Arumuganathane K, Ellisa MW, Mishler BD, Kelchf DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene 350: 117-128.

Wolf PG, Roper JM, Duffy AM (2010) The evolution of chloroplast genome structure in ferns. Genome 53: 731-738.

Wolf PG, Rowe CA, Hasebe M (2004) High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene 339: 89-97.

Wolf PG, Rowe CA, Sinclair RB, Hasebe M (2003) Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. DNA Research 10: 59-65.

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255.

Xu J-H, Liu Q, Hua W, Wang T, Xue Q, Messing J (2015) Dynamics of chloroplast genomes in green plants. Genomics 106: 221-231.

Zhu A, Guo W, Gupta S, Fan W, Mower JP (2016) Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytologist 209: 1747-1756.

Zhong B, Fong R, Collins LJ, McLenachan PA, Penny D (2014) Two new fern chloroplasts and decelerated evolution linked to the long generation time in tree ferns. Genome Biol Evol 6: 1166-1173.

Figure legends

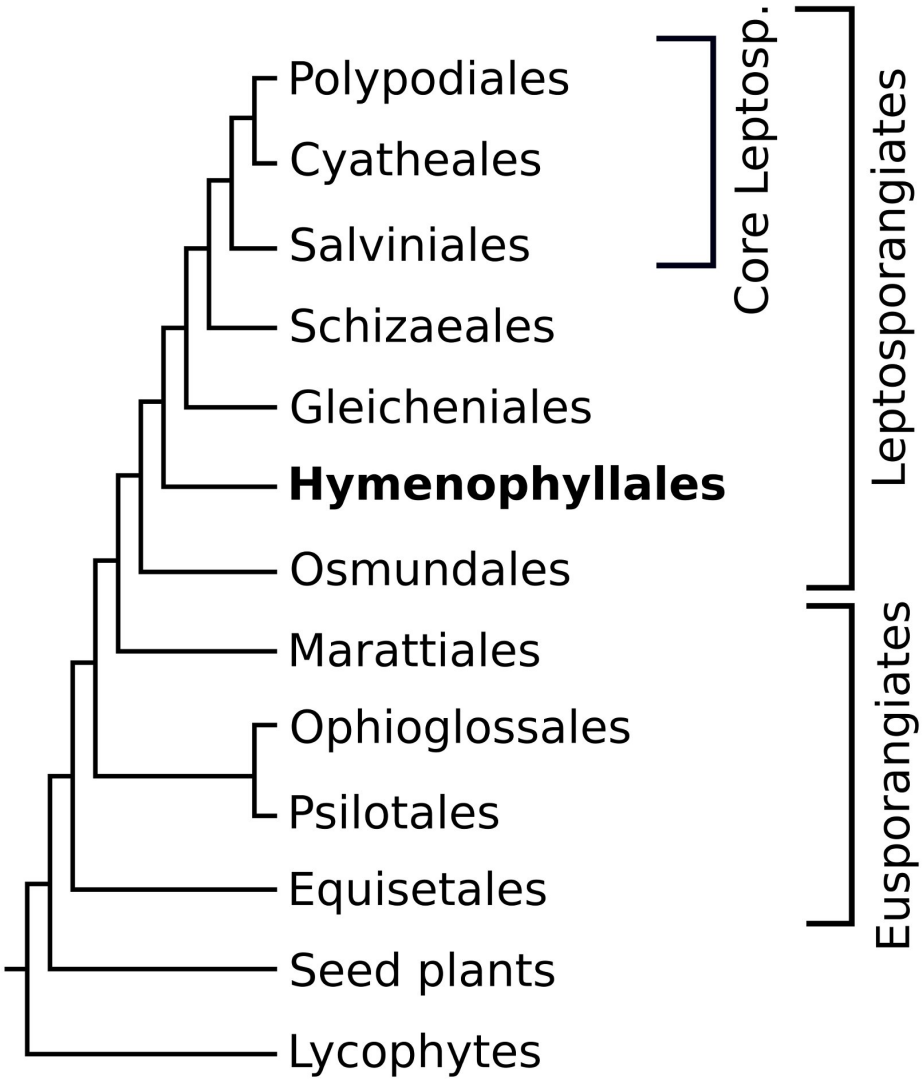

Figure 1. Schematic representation of currently accepted ordinal fern phylogeny (PPG I 2016) derived from the best available data (mainly from Schuettpelz and Pryer 2007; Rai and Graham 2010; Lehtonen 2011; Rothfels et al. 2015; Knie et al. 2015; Qi et al 2018).

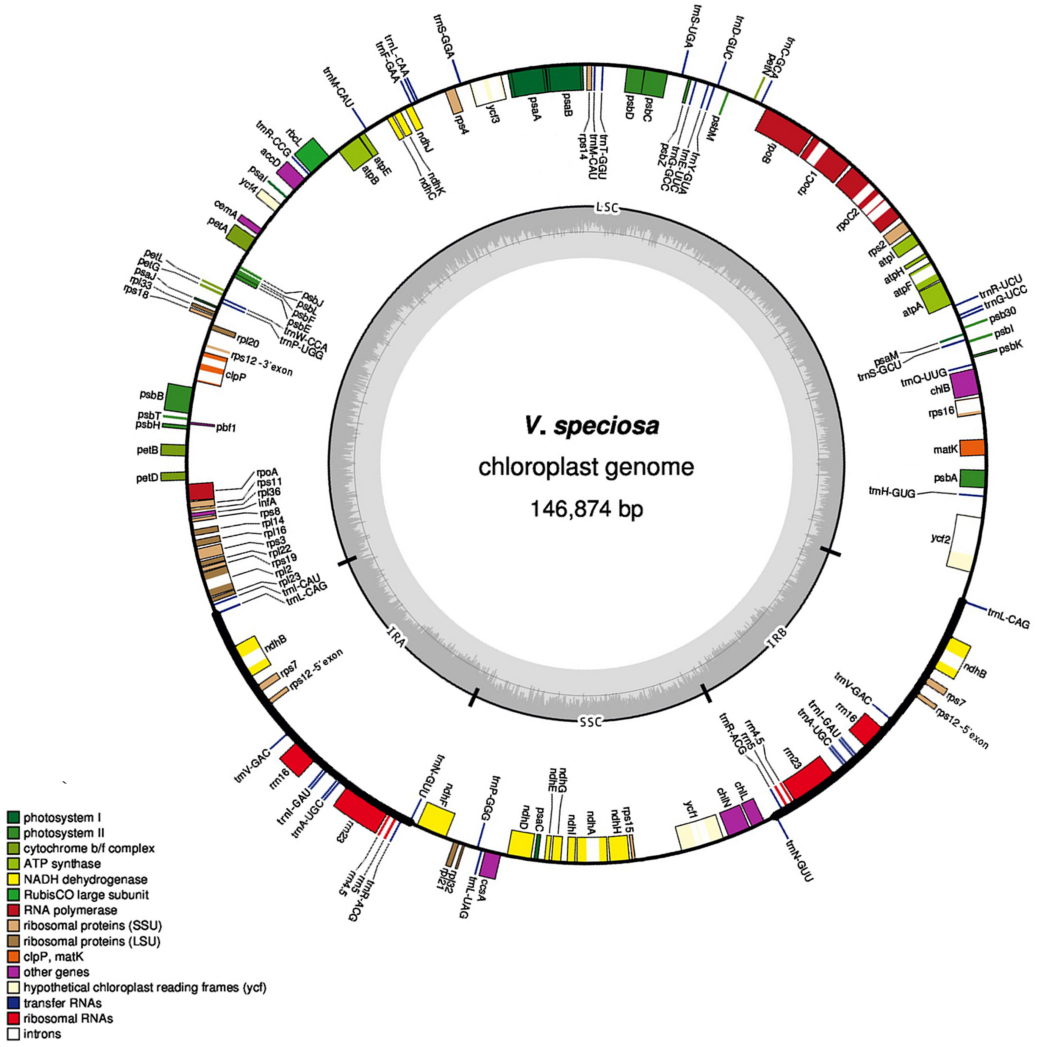

Figure 2. Gene map of the *V. speciosa* plastome. Boxes on the inside and outside of the outer circle represent genes transcribed clockwise and anti-clockwise, respectively. Gene boxes are color coded by functional groups as shown in the key. The inner circle displays the GC content represented by dark gray bars. The location of the IR and SC regions is marked on the inner circle. The location of the IRs is marked also on the outer cycle by thick black lines.

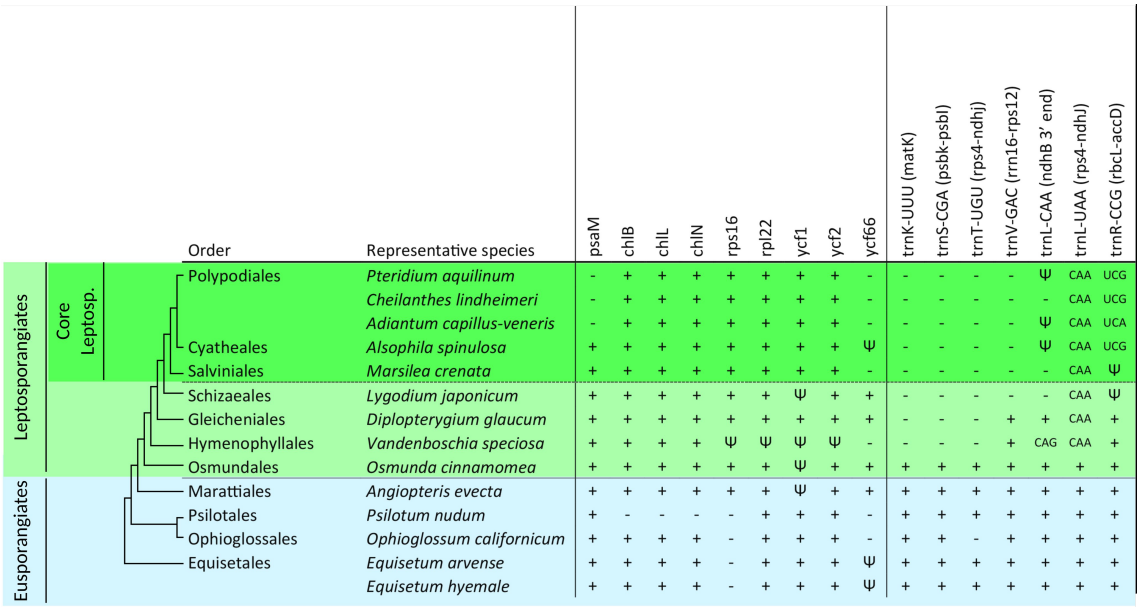

Figure 3. Presence (+) or absence (-) of different protein-coding and tRNA genes in the plastomes of representative species of each of the eleven fern orders. Ψ = pseudogene. Blue labeled lines include eusporangiate ferns while green lines include leptosporangiate ferns (darker for core leptospornagiate ferns).

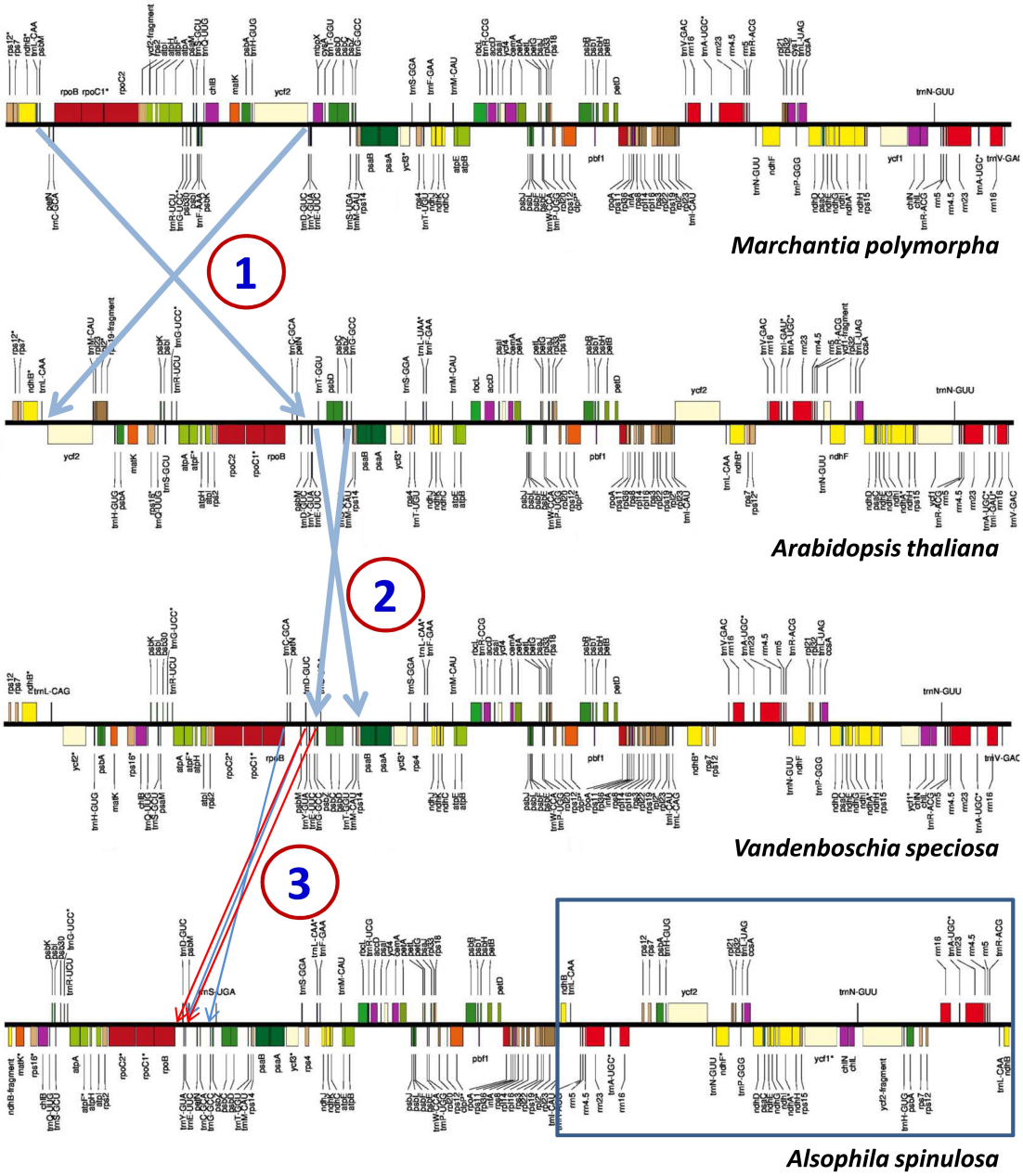

Figure 4. Linear gene map of *V. speciosa* plastome compared to those of a hornworth, *Marchantia polymorpha* (NC_001319.1), a seed plant, *Arabidopsis thaliana* (NC_000932.1), and a core leptosporangiate fern, *Alsophila spinulosa* (NC_012818.1). Numbers represent the inverted regions described in the text: between euphyllophytes and the rest of land plants (1); between ferns and the rest of land plants (2); and the inversion that characterizes core leptosportangiates (3). Blue arrows indicate the inversion ends. Red arrows indicate a change of location for the region comprising the *trnD-trnY-trnE* genes (that it is not inverted) as a consequence of the double inversion in the two region flanking it. A box stands out the rearranged IR regions in *Alsophila spinulosa* (the box also includes the LSC region). Genes with introns are labeled with asterisks.

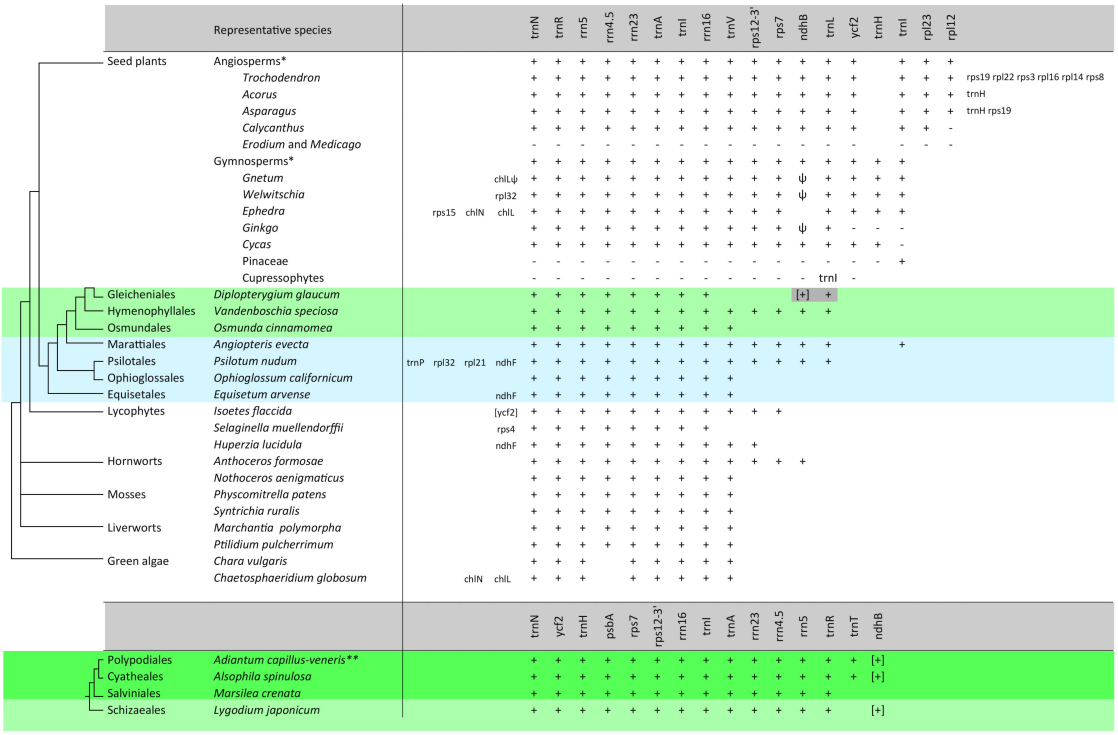

Figure 5. Inverted repeat (IR) gene content during land plant evolution. Rearranged IR regions of Core Leptosporangiates and Schizaeales are represented below. Shaded cells indicate an inverse order of these genes in this species due to an inversion. Genes in square brackets are those partially encoded in the IR. Pseudogenes are represented by Ψ. Blue labeled lines include eusporangiate ferns while green lines include leptosporangiate ferns (darker for core leptospornagiate ferns). *There are some specific losses/gains in specific lineages some of which have been represented below. **ndhB and trnT genes are pseudogenized in most species analyzed within this order, but not in *Adiantum capillus-veneris*.

Table 1. Main features and GenBank accession numbers of plastomes from different bryophyte, lycophyte, monylophyte and seed plants species.

		Accession number	Size (bp)	LSC (bp)	SSC (bp)	IRs (bp)	%GC	Genes	Protein-coding	tRNAs	rRNAs
Angiosperms	*Arabidopsis thaliana*	NC_000932.1	154478	84170	17780	26264	36.30	113	79	30	4
	Asparagus officinalis	NC_034777.1	156699	84999	18638	26531	37.76	112	78	30	4
	Glycine max	NC_007942.1	152218	83175	17895	25574	34.00	111	77	30	4
	*Erodium carvifolium**	NC_015083.1	116935	-	-	-	39.00	108	76	28	4
Gymnosperms	*Cycas taitungensis*	NC_009618.1	163403	90216	23039	25074	39.50	118	83	31	4
	Ginkgo biloba	NC_016986.1	156945	99221	22258	17733	39.60	120	81	35	4
	Welwitschia mirabilis	NC_010654.1	119726	68556	11156	20007	38.00	101	66	31	4
	*Pinus bungeana***	NC_028421.1	117861	65373	51538	475	38.80	111	71	36	4
Polypodiales	*Pteridium aquilinum*	NC_014348.1	152362	84335	21259	23384	41.53	117	84	29	4
	Adiantum capillus-veneris	NC_004766.1	150568	82282	21392	23447	42.01	117	84	29	4
Cyatheales	*Alsophila spinulosa*	NC_012818.1	156661	86308	21623	24365	40.43	117	85	28	4
Salviniales	*Marsilea crenata*	NC_022137.1	151628	87828	22210	20795	42.22	117	85	28	4
Schizaeales	*Lygodium japonicum*	NC_022136.1	157260	85448	21652	25080	40.64	118	86	28	4
Gleicheniales	*Diplopterygium glaucum*	NC_024158.1	151007	99857	21982	14584	40.16	119	84	30	4
Hymenophyllales	*Vandenboschia speciosa*	MH648610	146874	89620	21398	17928	37.50	120	85	31	4
Osmundales	*Osmunda cinnamomea*	NC_024157.1	142812	100294	22300	10109	40.23	121	84	33	4
Marattiales	*Angiopteris evecta*	NC_008829.1	153901	89709	22086	21053	35.48	121	85	32	4
Psilotales	*Psilotum nudum*	NC_003386.1	138829	84617	16304	18954	36.03	118	81	33	4
Ophioglossales	*Ophioglossum californicum*	NC_020147.1	138270	99058	19662	9775	42.15	120	84	32	4
Equisetales	*Equisetum arvense*	NC_014699.1	133309	93542	19469	10149	33.36	121	84	33	4
	Equisetum hyemale	NC_020146.1	131760	92580	18994	10093	33.74	121	84	33	4
Lycophytes	*Isoetes flaccida*	NC_014675.1	145303	91862	27205	13118	37.94	120	84	32	4
	Huperzia lucidula	NC_006861.1	154373	104088	19657	15314	36.30	119	86	29	4
Hornworts	*Anthoceros formosae*	NC_004543.1	161162	107503	22171	15744	32.90	120	84	32	4
Mosses	*Physcomitrella patens*	NC_005087.1	122890	85211	18501	9589	28.50	119	84	31	4
Liverworts	*Marchantia polymorpha*	NC_001319.1	121024	81095	19813	10058	28.80	124	89	31	4

*This species has lost the IRs
**This species has residual IRs

Table 2. Catalogue of genes identified in the plastome of *V. speciosa* classified according to gene class. Superscript numbers indicate the number of introns. Asterisk indicates that the gene is truncated at the 5' end.

Gene class	*Genes*					
ATP synthase	*atpA*	*atpB*	*atpE*	*atpF*[1]	*atpH*	*atpI*
Chlorophyll biosynthesis	*chlB*	*chlL*	*chlN*			
Cytochrome	*petA*	*petB*	*petD*	*petG*	*petL*	*petN*
Hypothetical protein	*ycf1*[1*]	*ycf2*[1*]	*ycf3*[2]	*ycf4*		
Miscellaneous proteins	*accD*	*cemA*	*ccsA*	*clpP*[2]	*infA*	*matK*
NADH dehydrogenase	*ndhA*[1]	*ndhB*[1]	*ndhC*	*ndhD*	*ndhE*	*ndhF*
	ndhG	*ndhH*	*ndhI*	*ndhJ*	*ndhK*	
Photosystem I	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*	*psaM*
Photosystem II	*psbA*	*psbB*	*psbC*	*psbD*	*psbE*	*psbF*
	psbH	*psbI*	*psbJ*	*psbK*	*psbL*	*psbM*
	pbf1(psbN)	*psbT*	*psbZ*	*psb30(ycf12)*		
RNA polymerase	*rpoA*	*rpoB*	*rpoC1*[1]	*rpoC2*[3]		
Rubisco	*rbcL*					
Ribosomal protein	*rpl2*[1]	*rpl14*	*rpl16*	*rpl20*	*rpl21*	*rpl22*[*]
	rpl23	*rpl32*	*rpl33*	*rpl36*		
	rps2	*rps3*	*rps4*	*rps7*	*rps8*	*rps11*
	rps12	*rps14*	*rps15*	*rps16*[1*]	*rps18*	*rps19*
tRNA genes	*trnA-UGC*[1]	*trnC-GCA*	*trnD-GUC*	*trnE-UUC*	*trnF-GAA*	*trnG-GCC*
	trnG-UCC[1]	*trnH-GUG*	*trnI-CAU*	*trnI-GAU*[1]	*trnL-CAA*[1]	*trnL-CAG*
	trnL-UAG	*trnfM-CAU*	*trnM-CAU*	*trnN-GUU*	*trnP-GGG*	*trnP-UGG*
	trnQ-UUG	*trnR-ACG*	*trnR-ACG*[1]	*trnR-CCG*	*trnR-UCU*	*trnS-GCU*
	trnS-GGA	*trnS-UGA*	*trnT-GGU*	*trnV-GAC*	*trnV-UAC*[1]	*trnW-CCA*
	trnY-GUA					
rRNA genes	*rrn23*	*rrn16*	*rrn5*	*rrn4.5*		

Table 3. List of genes with stop and non-canonical start codons.

Gene	Stop codon	Start codon
accD	CGA to UGA/CAA to UAA	
atpB	CAA to UAA	
atpH		ACG
atpI	CAA to UAA	
cemA	CAA to UAA	
chlB	CAA to UAA	
chlN	CAA to UAA	
clpP	UAU to UAA	
matK	CGA to UGA	
ndhA	CAA to UAA	
ndhF	CGA to UGA	
ndhG		ACG
ndhH		ACG
petA	CAA to UAA	
petD		AUA
rpoA	CAA to UAA	
rpoB	CGA to UGA and CAA to UAA	
rpoC1	CGA to UGA	
rpoC2	CGA to UGA and CAA to UAA	
rpl2	CGA to UGA	
rpl20	CAA to UAA	
rpl23	CGA to UGA	
rpl33	CGA to UGA	
rps2	CAA to UAA	
rps3	CGA to UGA	
rps4	CAA to UAA	
rps7	CGA to UGA	
rps8	CAA to UAA	
rps12	CAA to UAA	
rps15		AUA
rps18	CGA to UGA	
ycf4	CGA to UGA/CAA to UAA	

Table 4. Features and sequences of short direct and inverted repeats found in intergenic spacers of the *V. speciosa* plastome.

Direct repeats

Flanking genes	Size	Repeats	Identity (%)	Consensus
petN/psbM	6	12	70.8	ATTTAT
trnY-GUA/trnE-UUC	27	5	84.4	GAACGGATTTAGAGTCCGTCTCCCATC

Inverted repeats

Flanking genes	Size	Identity (%)	Repeat
ndhB/ trnL-CAG	25	88	GATTTACAAATTGACTCAGTAAAAG/CTAAATGTATAACTAAGTCATCTTC
trnL-CAA/trnF-GAA	27	100	CTGAAACGAACCCATTCAAGGAATCCA/GACTTTGCTTGGGTAAGTTCCTTAGGT
trnM-CAU/atpE	21	100	AAAACTTATTGGATGCCATAT/ATATGGCATCCAATAAGTTTT
rpl14/rpl16	19	100	ATAGGAAGAGATAGTTCCA/TATCCTTCTCTATCAAGGT
ycf1/chlN	22	100	TAGATTTATTAACCATAAAGGA/ATCTAAATAATTGGTATTTCCT