Transposable element landscapes illuminate past evolutionary events in the endangered fern *Vandenboschia speciosa*

Francisco J. Ruiz-Ruano[1,2,3], Beatriz Navarro-Domínguez [1,4], Juan Pedro M. Camacho [1] and Manuel A. Garrido-Ramos [1,*]

[1]Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain
[2]Department of Organismal Biology, Systematic Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[3]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom
[4]Department of Biology, University of Rochester, Rochester, New York 14627

*****Correspondence: mgarrido@ugr.es

Abstract

Vandenboschia speciosa is an endangered tetraploid fern species with a large genome (10.5 Gb). Its geographical distribution is characterized by disjoined tertiary flora refuges, with relict populations that survived past climate crises. Here we analyze the transposable elements (TEs) and found that they comprise about 76% of the *V. speciosa* genome, thus being the most abundant kind of DNA sequences in this gigantic genome. *V. speciosa* genome is composed of 51% and 5.6% of Class I and Class II elements, respectively. LTR retrotransposons were the most abundant TEs in this species (at least 42% of the genome), followed by non-LTR retrotransposons that constituted at least 8.7% of the genome of this species. We introduce an additional analysis to identify the nature of non-annotated elements (19% of the genome). A BLAST search of the non-annotated contigs against the *V. speciosa* TE database allowed determining the identity of almost half of them, which were most likely diverged sequence variants of the annotated TEs. In general, TE composition in *V. speciosa* resembles TE composition in seed plants. In addition, repeat landscapes revealed three episodes of amplification for all TEs, most likely due to demographic changes associated to past climate crises.

Keywords: climate crisis, demographic changes, endangered species, ferns, genome size, relict populations, tetraploidy, transposable elements, *Vandenboschia speciosa*.

Introduction

Transposable elements (TEs) are ubiquitous components of eukaryotic genomes that are considered drivers of genome evolution (Böhne et al. 2008; Belyayev 2014; Bourque et al. 2018), with relevant impact on both genome regulation (Slotkin and Martienssen 2007; Feschotte 2008; García-Pérez et al. 2016) and size (Gregory 2005; Bennetzen and Park 2018). A comprehensive analysis of how TE landscape contributes to a particular genome is thus relevant from a structural, functional and evolutionary perspective. Next-Generation Sequencing (NGS) and high-throughput *in silico* analysis of NGS reads have transformed the study of repetitive DNA, especially since the introduction of the RepeatExplorer (RE) pipeline which allows the identification and characterization of thousands of repetitive DNA elements on short NGS reads, by employing graph-based clustering of sequence reads (Novák et al. 2010, 2013, 2020b). Furthermore, the efficiency of repetitive DNA mining can be increased by means of the recursive application of the RE clustering algorithm combined with filtering out, at each round, the reads containing already known repetitive families (Ruiz-Ruano et al. 2016). Generated contigs are then properly annotated with appropriate software such as DANTE (http://repeatexplorer.org/), which tracks the REXdb database (Neumann et al. 2019).

As a general rule, there is a relationship between TE abundance and genome size, which contributes to explain the C-value paradox (Gregory 2005; Bennetzen and Park 2018). Indeed, it has been recently proved that, in land plants, genome size increases proportionally to repetitive DNA amount, reaching up to proportions of around 80% of repetitive DNA in large genomes (Novák et al. 2020a). Curiously, this trend is shifted in genomes larger than 10 Gb and the largest genomes might have about 55% of repetitive DNA, probably by the slow degradation of repeats over time (Novák et al. 2020a). Notwithstanding this, TE accumulation is not the only cause for genome size increase in plants, as polyploidization is considered to play a major role in plant genome size evolution (Alix et al. 2017; Vicient and Casacuberta 2017). In fact, it has been suggested that polyploidization might be the major factor contributing to the high chromosome numbers and large genomes in ferns (Klekowski and Baker 1966; Klekowski 1972; Wagner and Wagner 1980; Nakazato et al. 2008; Dyer et al. 2013; Marchant et al. 2019).

In this context, biological, life-history and genomic features, together with the phylogenetic position within vascular plants, make *Vandenboschia speciosa* an

attractive species for a genome-wide analysis with the aim to contribute to the knowledge of the impact of TEs in genome size and evolution in ferns. *V. speciosa* is a tetraploid fern species with a huge genome (10.5 Gb) (Manton 1950; Manton et al. 1986; Obermayer et al. 2002; Ebihara et al. 2007). This species is an endangered fern whose habitat is currently threatened by destruction and over-harvesting (Rumsey et al. 1999). It is a rare European-Macaronesian endemism, the only representative of a genus which has a primarily tropical distribution, with a current geographical distribution characterized by disjointed tertiary flora refuges in the European Atlantic coast and the Macaronesian islands (Canaries, Madeira and Azores), composed of relic populations with very few individuals that survived past climate crises (Rumsey et al. 1999). We have found that most DNA sequences in the genome of the fern *Vandenboschia speciosa* are TEs and that its specific TEs composition is similar to seed plants TEs composition. In addition, we analyzed repeat landscapes to investigate possible amplification events for each TE, in order to get insights on recent evolutionary pathways of these elements that could be important to understand the present relict distribution of this endangered species.

Materials and Methods

Materials

Vandenboschia speciosa sporophytes were collected at one out of seven populations located in the Alcornocales Natural Park (Cádiz, Spain): Canuto de Ojén-Quesada (OJEN). Sporophytes were frozen in liquid nitrogen in the field and stored at −80 °C. Genomic DNA (gDNA) was isolated using the DNeasy plant Mini kit (Quiagen). Pools of DNAs were generated from sets of five specimen DNAs and sequenced by Illumina HiSeq-2000 PE 2x101 nt technology, yielding about 16 Gb data (~1.5x coverage). Illumina sequencing data can be accessed at Sequence Read Archive (SRA) database under the BioProject PRJNA387541.

TE assembly and annotation

We performed an in-depth assembly of repetitive elements using RE (Novák et al. 2010, 2013, 2020b). For this, we first performed a quality trimming with Trimomatic (Bolger et al. 2014) to keep read pairs without adapters and a minimum quality of Q20. Then we randomly selected 2 x 2,000,000 Illumina reads with SeqTK

(https://github.com/lh3/seqtk) to run RE with default options. After one RE run, we extracted the most representative contigs for every cluster, specifically those representing up to a half of total cluster coverage with a custom script (https://github.com/fjruizruano/satminer/blob/master/rexp_get_contigs.py) and filtered out the reads from the original library that matches them using DeconSeq (Schmieder and Edwards 2011). Then, we randomly selected a new set of 2 x 2,000,000 reads from the filtered libraries, that were clustered with RE in a second round. Performing additional rounds of clustering and filtering had shown to be highly successful for satellite DNA (Ruiz-Ruano et al. 2016), as it allows detecting repetitive elements which, due to their low abundance, had gone unnoticed because their signals were masked by those of highly abundant elements. We annotated the resulting contigs by the DANTE software (http://repeatexplorer.org/) with the iterative search option and using as a reference the Viridiplantae v3.0 of REXdb (Neumann et al. 2019), i.e. a curated database for protein domains of plant repetitive. We considered separately the most conservative annotation in the "Final Classification" field. This classification is based on multiple top hits (the best hit + all other hits with score >= 80% of the score of the best hit). But sequences are classified on the deepest level showing no conflict among hits (Neumann et al. 2019). Thus, for example, a conflict between Class_I|LTR|Ty3/gypsy|chromovirus|Reina and Class_I|LTR|Ty3/gypsy|chromovirus|CRM, is resolved by DANTE as Class_I|LTR|Ty3/gypsy|chromovirus (Neumann et al. 2019). We annotated the contigs from the two RE rounds, excluding the "Simple_repeat" and "Low_complexity" contigs, and labeling the non-annotated contigs as "Unknown". Then, we used the msatcommander software (Faircloth 2008) to search for perfect microsatellite arrays (from 1 to 6 nt of monomer size) and removed arrays with 20 or more nucleotides. This is the minimum sensibility that RepeatMasker has to detect a microsatellite array. In addition, we screened the database with the CD-HIT program (Fu et al. 2012) using the options "-M 0 -aS 0.8 -c 0.8 -G 0 -g 1" in order to detect redundant contigs with at least a 80% of identity showing discrepant annotations. We did not find such kind of discrepancies in this sanity check. Finally, we combined all annotated RE contigs in a single database. As we were focused here on the study of TEs, we removed other repetitive elements from the final database, such as satDNA (Ruiz-Ruano et al. 2019a) and plastome sequences (Ruiz-Ruano et al. 2019b), which had previously been characterized in *V. speciosa*. In addition, we assembled the 45S and 5S ribosomal DNAs with the MITObim software (Hahn et al. 2013), using *Tetraplodon fuegianus* 45S (GenBank accession number KU095852) and *Marsilea*

quadrifolia 5S (GenBank accession number FR694363) as seeds. We then annotated the three types of elements with RepeatMasker (Smit et al. 2015) with "nolow" and "no_is" options and removed the contigs matching with these non-TE repetitive elements. Finally, we included the DANTE annotation for each RE cluster to the contigs IDs in the FASTA file with the RepeatMasker's format using a custom script (https://github.com/fjruizruano/ngs-protocols/blob/master/rexp_annot.py). The resulting database was deposited in FigShare (https://figshare.com/articles/dataset/Supplementary_Dataset_for_The_repeatome_of_the_endangered_fern_Vandenboschia_speciosa_/12124503).

Repeat landscapes

In order to estimate abundance and divergence for each annotated element, we aligned 5 million of randomly selected read pairs to the consensus sequences in the resulting RE database, using RepeatMasker with a custom script (https://github.com/fjruizruano/satminer/blob/master/repeat_masker_run_big.py). We used the calcDivergenceFromAlign.pl built-in tool of RepeatMasker to obtain a histogram of the Kimura 2-Parameter divergence for each element. Next, we transformed the abundance values to express them as genome proportion by dividing the number of aligned nucleotides by the total number of nucleotides in the selection of 10 million read pairs. The resulting histograms (hereafter referred to as Repeat Landscapes, RLs) were plotted in R.

Results

A first run of RE analysis allowed identifying 495 clusters of repetitive DNA sequences. However, an additional run of filtering+RE increased this figure up to 1,271 which were subsequently annotated by DANTE as TEs (Table 1). According to these annotations, TEs comprise at least 76% of the *V. speciosa* genome (Table 1).

As Table 1 shows, the *V. speciosa* genome has almost ten times the amount of retrotransposons related sequences (Class I elements) than DNA transposons related sequences (Class II elements), both kinds representing about 51% and 5.6% of the genome, respectively. By far, the most abundant sequences in *V. speciosa* are LTR retrotransposons (81.9% of Class I elements), belonging to two superfamilies (*Ty1/Copia* and *Ty3/Gypsy*), each representing at least about 18.5% of the genome

(Table 1). Only a percentage of *Ty1/Copia* (47% of *Copia* elements), could be assigned to a particular family, predominating *Ale* and *Tork* (Table 1). On the contrary, most *Ty3/Gypsy* sequences could be further annotated (99.3%), predominating the OTA group with *Athila* as the most representative element among *Ty3/Gypsy* elements (6.7% of the genome). *LINE* (Non-LTR retrotransposons TEs) sequences represent unusual amount in *V. speciosa* genome (at least 8.7% of the genome). Among DNA transposons, the higher amount of sequences belonged to *EnSpm-CACTA* elements, as it represents 62% of identified transposon sequences, followed by *Sola1* (36%) and *Helitron* (1.6%).

Almost 57% of the genome was annotated, thus remaining, at first instance, about 19% of the genome composed of "Unknown" elements (Table 1). In an effort to further characterize the non-annotated contigs, we blasted their sequences to the generated *V. speciosa* TE database, and found an important set of contigs that showed homology to some of the annotated sequences. Specifically, almost 32% of the "Unknown" sequences showed homology with annotated LTR elements (about 7% LTR/Copia and about 23% LTR/Gypsy), 4.6% with annotated LINEs and about 8% with DNA transposons (Table 2). Therefore, about 44% of non-annotated sequences could be somewhat identified by this procedure (Table 2). This allowed identifying about 8.5% (6.1% LTR, 0.9 LINE and 1.5 DNA transposons) of the *V. speciosa* genome as divergent variant sequences of TEs already annotated in Table 1. This raised the frequency of identified TEs till 65.4%, whereas the remaining 10.7% of TEs in the genome might be highly divergent or fern-specific TE sequences (Table 2).

It was remarkable to find that all TE superfamilies found within the genome of *V. speciosa* showed a similar profile for the Repeat Landscapes (RLs) built by comparing abundance and divergence of sequence variants (Figure 1). Thus, the landscapes are characterized by the presence of two to three well-defined peaks of abundance in most of the elements: one more diffuse representing sequences placed around 18% divergence, one peak around 13% divergence and the most conspicuous peak being around 4% of sequence divergence. It was also clear that this latter peak showed some slight differences among elements, as it was placed about 5% for LTR/Copia-Ale and LTR/Copia-Tork, 3% for DNA/Sola1 and LTR/Copia-GymcoII, 2% for LINE and Penelope as well as for LTR/Gypsy|chromovirus (see Figure 1).

Discussion

TEs largely contribute to V. speciosa genome size

To date, contrasting to other vascular plants, only the genomes of two heterosporous and one homosporous ferns have been sequenced (Sessa et al. 2014; Li et al. 2018; Marchant et al. 2019), and some other fragmentary data on TEs from a few fern genomes are available (Dyer et al. 2013; Wolf et al. 2015). However, because of their phylogenetic position, ferns are crucial for investigating TEs as well as other genomic traits. We wanted to contribute to this knowledge taking advantage of the development of new robust tools for the analysis of NGS reads. In this context, our present results revealed that 76% of the *V. speciosa* genome is composed of TEs, considerably improving our previous quantitative estimates of TEs in *V. speciosa* obtained after a single RE run (Ruiz-Ruano et al. 2019a). This is the highest proportion of TEs hitherto found in a fern genome (Wolf et al. 2015; Li et al. 2018; Marchant et al. 2019). Furthermore, our research confirm that TEs are the major component of the repeatome of *V. speciosa* while its tandem repetitive component comprised by satellite DNAs (about 0.4% of the genome) and microsatellites (about 2% of the genome) does not explain the huge genome size in this species (Ruiz-Ruano et al. 2019a).

The large fern genomes, especially homosporous ferns (average genome size 12 Gb; Sessa and Der 2016), are proposed to be paleopolyploid (reviewed in Barker 2013) behaving as diploid, and are characterized by extremely high numbers of chromosomes. As a result, polyploidization has been suggested as the major factor contributing to the high chromosome numbers and large genomes in ferns (Klekowski and Baker 1966; Klekowski 1972; Wagner and Wagner 1980; Nakazato et al. 2008; Dyer et al. 2013; Marchant et al. 2019). *V. speciosa* is considered to be a tetraploid species (Manton 1950; Manton et al. 1986; Obermayer et al. 2002), probably an allotetraploid (Ebihara et al. 2007), with 2n=144 chromosomes (Obermayer et al. 2002), which partly explains its large genome (1C= 10.52 Gb). However, we show here that TEs might be the main cause of genome size increase in this species, as they constitute 3/4 of genome sequences. In fact, recent papers claim that differences in fern genome size are attributable to TEs, and that fern repeat proportions are comparable to those of flowering plants (Li et al. 2018; Marchant et al. 2019). After analyzing genome size and spore size variation in the *Asplenium monanthes* fern complex, Dyer et al. (2013) concluded that other mechanisms, in addition to polyploidy, should explain genome size variation in ferns, and suggested "retrotransposon driven changes" as a possible cause. Our present results give support to this inference as retrotransposons actually constitute the immense majority of TEs

in *V. speciosa*. These data agree with the assumption that both TE transposition and polyploidization are considered major players in genome size evolution of plants (Alix et al. 2017; Vicient and Casacuberta 2017). In fact, Marchant et al. (2019) have recently found, in the model fern *Ceratopteris richardii* (11.25 Gb; n = 39), evidence suggesting that a single ancient polyploidy event and TE expansion both explain the large fern genomes, in resemblance to flowering plants. Furthermore, members of the fern order Salviniales (heterosporous ferns) that have smaller genome sizes than homosporous ferns also show differences in their repetitive content that explains some of the nearly threefold difference in genome size between *Salvinia* (*Salvinia cucullata*, 0.26 Gb; 25% of the genome are TEs) and *Azolla* (*Azolla filiculoides*, 0.75 Gb; 50% of the genome are TEs) (Li et al. 2018), suggesting that TE expansion appears to have been ubiquitous in ferns.

TE composition in V. speciosa resembles TE composition in seed plants

In contrast to animal genomes, LTR retrotransposons are the most abundant TEs in seed plant genomes (Wicker et al. 2007; López-Flores and Garrido-Ramos 2012). Likewise, LTR retrotransposons are the most abundant TE sequences in a few fern species analyzed up to date (Wolf et al. 2015; Li et al. 2018; Marchant et al. 2019), and they represent about 42% of *V. speciosa* genome (48% if we take into account the divergent elements identified by BLAST) (Tables 1 and 2). In order to contribute to a better understanding of fern TEs, we identified some familiar ascription among the *Ty1/Copia* and *Ty3/Gypsy* TEs. Half of the *LTR/Copia* sequences that we found (53%) were classified as generic Ty1/Copia elements. However, the other 47% belonged specifically to five families (Ale, Tork, GymcoII, GymcoIII and GymcoIV), which are usually present in seed plants but are absent in non-vascular plants (Bryophyta) and Lycopodiophyta (Neumann et al. 2019). Specifically, Gymco elements (I to IV) are specific of gymnosperms, whereas Ale and Tork are common to gymnosperms and flowering plants (Neumann et al. 2019). We did not detect *LTR/Copia-Bryco* or *LTR/Copia-Lyco* elements, which are the only families found in Bryophyta and Lycopodiophyta, respectively. Among Ty3/Gypsy elements, chromoviruses represent the oldest and most widespread lineage of Ty3/Gypsy retrotransposons in seed plants (Novikov 2012; Neumann et al. 2019). In consistency, three of these families (*CRM*, *Reina* and *Galadriel*) were present in the genome of *V. speciosa*, the latter being also found in Lycophyta. Notwithstanding, they represent only about 12% of *LTR/Gypsy* elements (Table 1). Among the non-chromoviruses (87% of *LTR/Gypsy* elements), OTA

were the only type found in this genome (Table 1). Many of the OTA sequences detected (55.7%) could not be further annotated. Among the remaining OTA elements, *Athila* was the most represented *LTR/Gypsy* element in *V. speciosa* (6.7% of the genome; almost the 9% of the genome if we consider the BLAST analysis), followed by Tat, both being typical of vascular plants, also found in lycophytes (Neumann et al. 2019). Remarkably, we did not find *Phygy* elements, which are specific of Bryophyta, or *Selgy* elements which are specific of Lycopodiophyta (Neumann et al. 2019). These results suggest that ferns share more classes of LTR elements with seed plants than with other basal groups of plant phylogeny, whether vascular (Lycophyta) or non-vascular (Bryophyta), in concordance with current phylogenies of vascular plants (Pryer et al. 2001; Schneider et al. 2004; Smith et al. 2006).

LINE retrotransposons comprised about 8.7% of the *V. speciosa* genome (9.6% considering the BLAST results), a very high figure compared with other plant genomes (average < 1%) (Hřibová et al. 2010; Novikov et al. 2012; Makałowski et al. 2019). This finding was previously pointed out by Wolf et al. (2015), although they estimated lower values (average= 2.2%) in the fern species analyzed. Interestingly, *Penelope* represents 0.54% of the *V. speciosa* genome, whereas it is rarely identified in plant genomes despite its wide distribution among eukaryotes, including the spike moss *Selaginella moellendor* (Arkhipova 2006; Novikov et al. 2012; Tollis and Boissinot 2012).

Finally, DNA transposons constitute about 1%-15% of plant genomes (Novikov et al. 2012; Weiss-Schneeweiss et al. 2015), and their proportion in the genome of *V. speciosa* (5.6-7.1%) was within this range, with *CACTA* and *Sola1* as predominating elements, as in other fern genomes (Li et al. 2018). Interestingly, about 1.6% of all annotated DNA transposons in *V. speciosa* belong to the order *Helitron*, a kind of rolling-circle transposons that have demonstrated a tremendous potential for gene shuffling and duplication in plants (Morgante et al. 2005; Thomas and Pritham 2015).

Taken together, the TE composition found in the genome of the fern *V. speciosa* shows high resemblance with seed plants, especially in the case of LTR retrotransposons.

BLAST search of the V. speciosa database allowed the increase of the proportion of identified TEs

Successful TEs annotation depends on the similarity of the TEs found in the studied genome with those available in TE databases, which currently are biased toward model organisms which, in the present case, are phylogenetically distant species. In addition, the sequence of inactive TEs diverges through mutation and drift. Thus, the particular TE landscape in each species is composed of a number of repeats that rapidly diverge both at the intra- and inter-specific levels and this makes it difficult to properly identifying genome-specific sequence variants for each element (Neumann et al. 2019). Therefore, it is conceivable that most of the 19% of the genome containing non-annotated TEs in *V. speciosa* is made up of diverged TE sequences. In this respect, we further characterized the non-annotated contigs obtained with RE using a BLAST search of the *V. speciosa* TE database. Overall, we identified by this procedure the nature of an additional 8.44% proportion of the genome. All together, DANTE annotation and BLAST identification of the non-annotated sequences revealed that TEs represent about 65.3% of the *V. speciosa* genome, whereas another 10.7% of TEs consists of unidentified TEs, which most likely were too divergent to be identified with the methods employed here. Anyway, we cannot rule out that some of these unidentified sequences could correspond to fern-specific (even functional) TEs.

Temporal changes in TE abundance

RLs showed one prominent and two less pronounced peaks of TE abundance relative to sequence divergence (see Figure 1). These peaks represent conspicuous sets of repeats grouped around specific values of sequence divergence (i.e. 4%, 13% and 18%). As sequence divergence is due to mutational changes and these are proportional to time, we infer that these three peaks are indicative of temporally different TE expansion waves within the genome of this species. The fact that repeat landscape profiles were highly similar for all TEs, we infer that these expansion waves were associated with demographic changes in the ancestral populations of the two analyzed here. Current localities of *V. speciosa* are small disjoined tertiary flora refuges harboring relic populations that survived the glacial cycles. Several important climatic change events during the last 5 my (such as the Messinian Salinity Crisis, the Pliocene-Pleistocene transition with the establishment of Mediterranean climate and extinction of typical tertiary taxa, and the Pleistocene with interglacial cycles) might have influenced evolutionary pathways in *V. speciosa* resulting from successive contractions of the area of distribution of the species, population fragmentation and

isolation leading to bottlenecks eroding genetic variability through genetic drift (Ben-Menni Schuler 2019). Previous research indicated that reduced effective population size can trigger an increase in TE copy number and genome size (Lynch and Conery 2003; García-Guerreiro 2012; Bourgeois and Boissinot 2019). According to these authors, while most new TE insertions would be eliminated by selection in large populations, drift would predominate over selection in small populations and thus TE abundance could eventually increase. It is thus conceivable that successive bottlenecks in *V. speciosa* could have boosted the massive TE expansions reported here. Similar increases in TE copy numbers in small populations after bottlenecks have also been found in *Arabidopsis lyrata* (Lockton et al. 2008; Ross-Ibarra et al. 2008) and *Drosophila subobscura* (García-Guerreiro et al. 2008) as a consequence of strong effect of stochastic events and a reduced efficiency of purifying selection in those populations (reviewed in Bourgeois and Boissinot 2019). Interestingly, it cannot be ruled out that the environmental stresses associated to the mentioned events might also be important factors associated to TE activation (Capy et al. 2000; Kalendar et al. 2000; García-Guerreiro 2012; Chuong et al. 2017; Bourgeois and Boissinot 2019).

Acknowledgments

This research has been financed by the Spanish Ministerio de Economía y Competitividad and FEDER founds, grant: CGL2010-14856 (subprograma BOS). FJ Ruiz-Ruano was supported by a Junta de Andalucía fellowship (Spain), a postdoctoral fellowship from Sven och Lilly Lawskis fond (Sweden) and a Marie Skłodowska-Curie Individual Fellowship grant agreement 875732 (EU). 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 the team of Agentes de Medio Ambiente of the Consejería, helped us with the sampling procedure.

References

Alix, K., Gérard, P.R., Schwarzacher, T., Heslop-Harrison, J.S. (Pat). 2017. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Ann. Bot., **120**:183-194.

Arkhipova, I.R. 2006. Distribution and phylogeny of Penelope-like elements in eukaryotes. Syst. Biol., **55**: 875-885.

436 Barker, M.S. 2013. Karyotype and genome evolution in Pteridophytes. In:
437 Leitch I.J. et al. (eds) Plant Genome Diversity. Volume 2, Physical Structure,
438 Behaviour and Evolution of Plant Genomes. Springer-Verlag, Wien, pp 245-253.
439 Belyayev, A. 2014. Bursts of transposable elements as an evolutionary driving
440 force. J. Evol. Biol., **27**: 2573-2584.
441 Ben-Menni Schuler, S. 2019. Genetic diversity and phylogeographic pattern of
442 four Tertiary-relict fern species: *Vandenboschia speciosa* (Willd.) G. Kunkel,
443 *Diplazium caudatum* (Cav.) Jermy, *Pteris incompleta* Cav. and *Culcita macrocarpa* C.
444 Presl. PhD Thesis, University of Granada, Granada, Spain.
445 http://hdl.handle.net/10481/58616
446 Bennetzen, J.L., Park, M. 2018. Distinguishing friends, foes, and freeloaders
447 in giant genomes. Curr. Opin. Genet. Dev., **49**: 49-55.
448 Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., Volff, J.N. 2008.
449 Transposable elements as drivers of genomic and biological diversity in vertebrates.
450 Chromosome Res., **16**: 203-215.
451 Bolger, A.M., Lohse, M., Usadel, B. 2014. Trimmomatic: a flexible trimmer for
452 Illumina sequence data. Bioinformatics, **30**: 2114-2120.
453 Bourgeois, Y., Boissinot, S. 2019. On the population dynamics of junk: a
454 review on the population genomics of transposable elements. Genes, **10**: 419.
455 Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell,
456 M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C.
457 2018. Ten things you should know about transposable elements. Genome Biology, **19**:
458 199.
459 Capy, P., Gasperi, G., Biémont, C., Bazin, C. 2000. Stress and transposable
460 elements: coevolution or useful parasites. Heredity, **85**: 101-106.
461 Chuong, E.B., Elde, N.C., Feschotte, C. 2017. Regulatory activities of
462 transposable elements: From conflicts to benefits. Nature Review Genetics,**18**: 71-86.
463 Dyer, T.J., Pellicer, J., Savolainen, V., Leitch, I.J., Schneider, H. 2013.
464 Genome size expansion and the relationship between nuclear DNA content and spore
465 size in the *Asplenium monanthes* fern complex (Aspleniaceae). BMC Plant Biology, **13**:
466 219.
467 Ebihara, A., Dubuisson, J.Y., Iwatsuki, K., Ito, M. 2007a. Systematics of
468 *Trichomanes* (Hymenophyllaceae: Pteridophyta), progress and future interests. Fern
469 Gazette, **18**: 53-58.

Faircloth, B.C. 2008. MSATCOMMANDER: Detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources, **8**: 92-94.

Feschotte, C. 2008. The contribution of transposable elements to the evolution of regulatory networks. Nat. Rev. Genet., **9**: 397-405.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, **28**: 3150-3152.

García-Guerreiro, M.P. 2012. What makes transposable elements move in the *Drosophila* genome? Heredity, **108**: 461-468.

García-Guerreiro, M.P., Chávez-Sandoval, B.E., Balanyà, J., Serra, L., Fontdevila, A. 2008. Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of *Drosophila subobscura*. BMC Evol. Biol. **8**: 234.

García-Pérez, J.L., Widmann, T.J., Adams, I.R. 2016. The impact of transposable elements on mammalian development. Development, **143**: 4101-4114.

Gregory, T.R. 2005. Genome size evolution in animals. In: Gregory T.R. (ed.) The Evolution of the Genome. Elsevier Academic Press, Burlington, USA, pp 3-87.

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.

Hřibová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J., Dolezel, J. 2010. Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biology, **10**: 204.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Nat. Acad. Sci. USA **97**: 6603-6607.

Klekowski, E. 1972. Genetical features of ferns as contrasted with seed plants. Ann. Missouri Bot. Gard. **59**: 138-151.

Klekowski, E.J., Baker, H.G. 1966. Evolutionary significance of polyploidy in the Pteridophyta. Science, **153**: 305-307.

Li, F.W., et al. 2018. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nature Plants, **4**: 460-472.

Lockton, S., Ross-Ibarra, J., Gaut, B.S. 2008. Demography and weak selection drive patterns of transposable element diversity in natural populations of *Arabidopsis lyrata*. Proc. Nat. Acad. Sci. USA **105**: 13965-13970.

López-Flores, I., Garrido-Ramos, M.A. (2012) The repetitive DNA content of eukaryotic genomes. In: Garrido-Ramos, M.A. (ed.) Repetitive DNA. Genome Dynamics Volume 7. Karger, Basel, Switzerland, pp 126-152.

Lynch, M., Conery, J.S. 2003. The origins of genome complexity. Science, **302**: 1401-1404.

Makałowski, W., Gotea, V., Pande, A., Makałowska, I. 2019. Transposable Elements: classification, identification, and their use as a tool for comparative genomics. In: Anisimova M (ed) Evolutionary Genomics: Statistical and Computational Methods, Methods in Molecular Biology Volume 1910. Springer Nature, Switzerland, pp 177-207.

Manton, I. 1950. Problems in cytology and evolution in the Pteridophyta. Cambridge University Press, Cambridge, 316 pp.

Manton, I., Lovis, J.D., Vida, G., Gibby, M. 1986. Cytology of the fern flora of Madeira. Bulletin of the British Museum (Natural History), Botany **15**: 123-161.

Marchant, D.B., Sessa, E.B., Wolf, P.G., Heo, K., Barbazuk, W.B., Soltis, P.S., Soltis, D.E. 2019. The C-Fern (*Ceratopteris richardii*) genome: insights into plant genome evolution with the first partial homosporous fern genome assembly. Sci. Rep., **9**: 18181.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., Rafalski, A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nature Genetics, **37**: 997-1002.

Nakazato, T., Barker, M.S., Rieseberg, L.H., Gastony, G.J. 2008. Evolution of the nuclear genome of ferns and lycophytes. In: Ranker TA, Haufler CH (eds) Biology and Evolution of Ferns and Lycpohytes. Cambridge University Press, Cambridge, pp. 175-198.

Neumann, P., Novák, P., Hoštáková, N., Macas, J. 2019. Systematic survey of plant LTR retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA, **10**: 1.

Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblizkova, A., Dodsworth, S., Kovarik, A., Pellicer, J., Wang, W., Macas, J., Leitch, I.J., Leitch, A.R. 2020a. Repeat-sequence turnover shifts fundamentally in species with large genomes. Nature Plants, **6**: 1325-1329.

Novák, P., Neumann, P., Macas, J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics **11**: 378.

Novák, P., Neumann, P., Macas, J. 2020b. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nature Protocols **15**: 3745-3776.

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.

Novikov, A., Smyshlyaev, G., Novikova, O. 2012. Evolutionary history of LTR retrotransposon chromodomains in plants. International Journal of Plant Genomics, **2012**: 874743.

Obermayer, R., Leitch, I.J., Hanson, L., Bennett, M.D. 2002. Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. Ann. Bot. **90**: 209-217.

Pryer, K.M., Schneider, H., Smith, A.R., Cranfill, R., Wolf, P.G., Hunt, J.S., Sipes, S.D. 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature, **409**: 618-622.

Ross-Ibarra, J., Wright, S.I., Foxe, J.P., Kawabe, A., DeRose-Wilson, L., Gos, G., Charlesworth, D., Gaut, B.S. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS ONE 2008: 3.

Ruiz-Ruano, F.J., López-León, M.D., Cabrero, J., Camacho, J.P.M. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci. Rep. **6**: 28333.

Ruiz-Ruano, F.J., Navarro-Domínguez, B., Camacho, J.P.M., Garrido-Ramos, M.A. 2019a. Characterization of the satellitome in lower vascular plants: the case of the endangered fern *Vandenboschia speciosa*. Ann. Bot. **123**: 587-599.

Ruiz-Ruano, F.J., Navarro-Domínguez, B., Camacho, J.P.M., Garrido-Ramos, M.A. 2019b. Full plastome sequence of the fern *Vandenboschia speciosa* (Hymenophyllales): structural singularities and evolutionary insights. J. Plant. Res. **132**: 3-17.

Rumsey, F.J., Vogel, J.C., Russell, S.J., Barrett, J.A., Gibby, M. 1999. Population genetics and conservation biology of the endangered fern *Trichomanes speciosum* (Hymenophyllaceae) in Scotland. Biol. J. Linn. Soc. **66**: 333-344.

Schmieder, R., Edwards, R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE **6**: e17288.

Schneider, H., Schuettpelz, E., Pryer, K.M., Cranfill, R., Magallón, S., Lupia, R. 2004. Ferns diversified in the shadow of angiosperms. Nature, **428**: 553-557.

576 Sessa, E.B., et al. 2014. Between Two Fern Genomes. Gigascience **2014**: 3.
577 Sessa, E.B., Der, J.P. 2016. Evolutionary genomics of ferns and lycophytes. In:
578 Advances in Botanical Research 78, Elsevier, pp. 215-254.
579 Slotkin, R.K., Martienssen, R. 2007. Transposable elements and the epigenetic
580 regulation of the genome. Nat. Rev. Genet. **8**: 272-285.
581 Smit, A.F.A., Hubley, R., Green, P. 2015. RepeatMasker Open-4.0. 2013-2015.
582 http://repeatmasker.org.
583 Smith, A.R., Pryer, K.M., Schuettpelz, E., Korall, P., Schneider, H., Wolf, P.G.
584 2006. A classification for extant ferns. Taxon **55**: 705-731.
585 Thomas, J., Pritham, E.J. 2015. Helitrons, the eukaryotic rolling-circle
586 transposable elements. Microbiology Spectrum **3**: MDNA3-0049-2014.
587 Tollis, M., Boissinot, S. 2012. The evolutionary dynamics of transposable
588 elements in eukaryote genomes. In: Garrido-Ramos, M.A. (ed.) Repetitive DNA.
589 Genome Dynamics Volume 7. Karger, Basel, Switzerland, pp 68-97.
590 Vicient, C.M., Casacuberta, J.M. 2017. Impact of transposable elements on
591 polyploid plant genomes. Ann. Bot. **120**: 195-207.
592 Wagner, W.H., Wagner, F.S. 1980. Polyploidy in pteridophytes. In: Polyploidy.
593 Springer, pp 199-214.
594 Weiss-Schneeweiss, H., Leitch, A.R., McCann, J., Jang, T.S., Macas, J. 2015.
595 Employing next generation sequencing to explore the repeat landscape of the plant
596 genome. In: Hörandl, E., Appelhans, M. (eds.) Next Generation Sequencing in Plant
597 Systematics Regnum Vegetabile. Koeltz Scientific Books, Königstein, Germany, pp
598 155-179.
599 Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B.,
600 Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman,
601 A.H. 2007. A unified classification system for eukaryotic transposable elements.
602 Nature Rev. Gen. **8**: 973-982.
603 Wolf, P.G., Sessa, E.B., Marchant, D.B., Li, F.W., Rothfels, C.J., Sigel, E.M.,
604 Gitzendanner, M.A., Visger, C.J., Banks, J.A., Soltis, D.E., Soltis, P.S., Pryer, K.M.,
605 Der, J.P. 2015. An exploration into fern genome space. Genome Biol. Evol., **7**: 2533-
606 2544.
607
608
609
610
611

Figure legends

Figure 1. Curve profiles in the Repeat Landscapes (RL) of TEs for *V. speciosa*

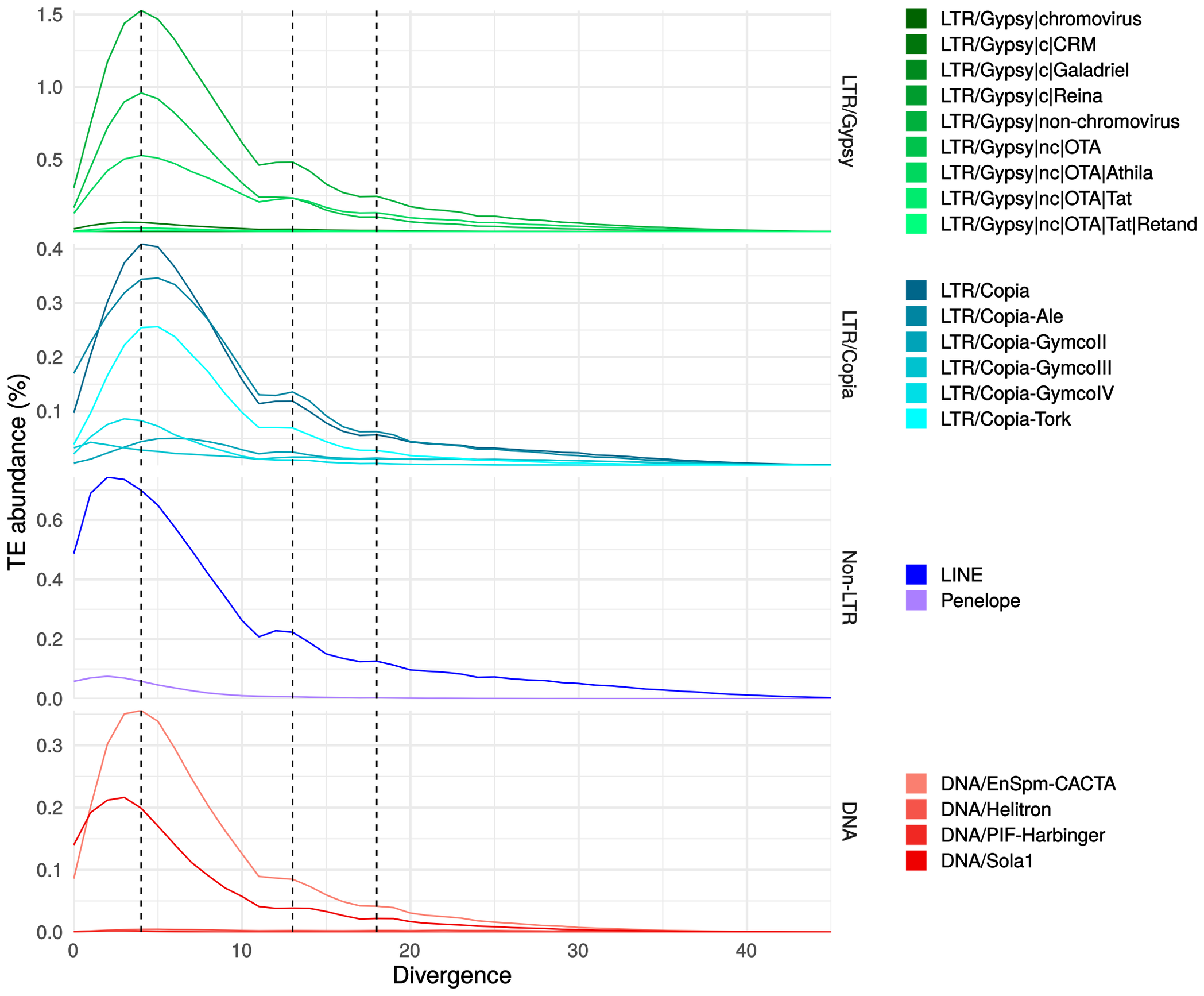
Table 1. TE content of *V. speciosa* genome.

TE taxonomy				Abundance	Abundance in respect to all TEs	Abundance in respect to annotated TEs
Class	**Order**	**Superfamily**	**Family**			
Class I	LTR			4,96%	6,53%	8,72%
		Ty1/Copia		9,90%	13,02%	17,40%
			LTR/Copia-Ale	4,27%	5,62%	7,50%
			LTR/Copia-Tork	2,43%	3,20%	4,27%
			LTR/Copia-GymcoII	0,70%	0,92%	1,23%
			LTR/Copia-GymcoIV	0,65%	0,86%	1,14%
			LTR/Copia-GymcoIII	0,58%	0,76%	1,02%
			TOTAL COPIA	**18,53%**	**24,38%**	**32,57%**
		Ty3/Gypsy		0,13%	0,17%	0,23%
		Non-chromovirus	LTR/Gypsy-OTA	9,04%	11,89%	15,89%
			LTR/Gypsy-OTA|Athila	6,73%	8,85%	11,83%
			LTR/Gypsy-OTA|Tat	0,32%	0,42%	0,56%
			LTR/Gypsy-OTA|Tat|Retand	0,14%	0,18%	0,25%
			TOTAL GYPSY NON-CHROMO	**16,23%**	**21,35%**	**28,52%**
		Chromovirus		1,44%	1,89%	2,53%
			LTR/Gypsy-CRM	0,69%	0,91%	1,21%
			LTR/Gypsy-Reina	0,06%	0,08%	0,11%
			LTR/Gypsy-Galadriel	0,02%	0,03%	0,04%
			TOTAL GYPSY CHROMO	**2,21%**	**2,91%**	**3,88%**
			TOTAL GYPSY	**18,57%**	**24,43%**	**32,64%**
	Total LTR			**42,06%**	**55,33%**	**73,92%**
	LINE			**8,71%**	**11,46%**	**15,31%**
	PLE	Penelope		**0,54%**	**0,71%**	**0,95%**
	Caulimovirus			**0,03%**	**0,04%**	**0,05%**
Total Class I				**51,34%**	**67,54%**	**90,23%**
Class II						
Subclass I	EnSpm-CACTA			3,46%	4,55%	6,08%
	Sola1			2,00%	2,63%	3,51%
	PIF-Harbinger			0,01%	0,01%	0,02%
	Total			**5,47%**	**7,20%**	**9,61%**
Subclass II	**Helitron**			**0,09%**	**0,12%**	**0,16%**
Total Class II				**5,56%**	**7,31%**	**9,77%**
Total annotated elements				**56,90%**	**74,86%**	**100,00%**
Unknown				19,11%	25,14%	
TOTAL				**76,01%**	**100,00%**	

Table 2. TE identification among non-annotated TEs of the genome of *V. speciosa*. Abundance: percentage in the genome of each non-annotated element but identified as a specific kind of TE by BLAST; Abundance in respect to Unknown elements: percentage of each identified TE in respect to the total of non annotated elements; Total identified elements: percentage of each TE in the genome of *V. speciosa* taken together both annotation and BLAST identification.

TE taxonomy			Abundance	Abundance in respect to Unknown elements	Total identified elements (Annotation + Blast identification)
Class	**Order**	**Superfamily**			
Class I	LTR		0,38%	1,99%	5,34%
		Ty1/Copia*	1,31%	6,86%	19,84%
		Ty3/Gypsy**	4,38%	22,92%	22,95%
	Total LTR		**6,07%**	**31,76%**	**48,13%**
	LINE		**0,87%**	**4,55%**	**9,58%**
	PLE	Penelope	**0,00%**	**0,00%**	**0,54%**
	Caulimovirus		**0,00%**	**0,00%**	**0,03%**
Total Class I			**6,94%**	**36,32%**	**58,28%**
Class II					
Subclass I	EnSpm-CACTA		0,46%	2,41%	3,92%
	Sola1		1,04%	5,44%	3,04%
	PIF-Harbinger		0,00%	0,00%	0,01%
	Total		**1,50%**	**7,85%**	**6,97%**
Subclass II	**Helitron**		**0,00%**	**0,00%**	**0,09%**
Total Class II			**1,50%**	**7,85%**	**7,06%**
Total annotated elements			**8,44%**	**44,17%**	**65,34%**
Unknown			10,67%	55,83%	10,67%
TOTAL			**19,11%**	**100,00%**	**76,01%**

*(0.22% Tork, 0.06% Ale, 0,04 GymcoII, 0.03% GymcoIII and 0.06% GymcoIII)
**(4.24% OTA|Athila, 1.86% OTA, 0.15 OTA|Tat|Retand, 0.13 Chromovirus)


TE abundance (%)
Divergence
LTR/Gypsy
LTR/Copia
Non-LTR
DNA
LTR/Gypsy|chromovirus
LTR/Gypsy|c|CRM
LTR/Gypsy|c|Galadriel
LTR/Gypsy|c|Reina
LTR/Gypsy|non-chromovirus
LTR/Gypsy|nc|OTA
LTR/Gypsy|nc|OTA|Athila
LTR/Gypsy|nc|OTA|Tat
LTR/Gypsy|nc|OTA|Tat|Retand
LTR/Copia
LTR/Copia-Ale
LTR/Copia-GymcoII
LTR/Copia-GymcoIII
LTR/Copia-GymcoIV
LTR/Copia-Tork
LINE
Penelope
DNA/EnSpm-CACTA
DNA/Helitron
DNA/PIF-Harbinger
DNA/Sola1