J Mol Evol (2008) 66:405–415
DOI 10.1007/s00239-008-9095-3

Multilevel Control of Organelle DNA Sequence Length in Plants

Jérôme Duminil · Delphine Grivet · Sébastien Ollier · Sylvain Jeandroz · Rémy J. Petit

Received: 29 October 2007 / Accepted: 25 February 2008 / Published online: 1 April 2008

Abstract We have compared the length of noncoding organelle DNA spacers in a broad sample of plant species characterized by different life history traits to test hypotheses regarding the nature of the mechanisms driving changes in their size. We first demonstrate that the spacers do not evolve at random in size but have experienced directional evolutionary trends during plant diversification. We then study the relationships between spacer lengths and other molecular features and various species attributes by taking into account population genetic processes acting within cell lineages. Comparative techniques are used to test these relationships while controlling for species phylogenetic relatedness. The results indicate that spacer length depends on mode of organelle transmission, on population genetic structure, on nucleotide content, on rates of molecular evolution, and on life history traits, in conformity with predictions based on a model of intracellular competition among replicating organelle genomes.

Keywords Organelle inheritance · Chloroplast · G_{ST} · Mitochondria · Nucleotide composition · Replication race

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9095-3) contains supplementary material, which is available to authorized users.

J. Duminil · D. Grivet · R. J. Petit (✉)
INRA, UMR Biodiversité, Gènes & Communautés, 69 route d'Arcachon, Cestas 33612, France
e-mail: petit@pierroton.inra.fr

S. Ollier
Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, UMR 5558, Lyon, France

S. Jeandroz
Interactions Arbres/Micro-organismes, Université Nancy I, UMR 1136, Nancy, France

Present Address:
J. Duminil
Laboratoire d'Eco-Ethologie Evolutive, Université Libre de Bruxelles, Bruxelles, Belgium

Present Address:
D. Grivet
Department of Forest Systems and Resources, Forest Research Institute, CIFOR-INIA, Madrid 28040, Spain

Present Address:
S. Jeandroz
UPSP PROXISS, ENESAD, Dijon, France

Introduction

Evolutionary forces such as selection, migration, and drift take place at multiple hierarchical levels of biological organization and can feed back to lower levels (Rand 2001). As a consequence, a number of genome features should depend on population or species attributes. Among the numerous and often unexpected examples reported so far, one can cite an increased frequency of introns as a "pathological" response to small population size (Lynch 2002), an increase in the average allele size of microsatellites in geographically peripheral maize groups (Vigouroux et al. 2003), faster rates of substitution for nonsynonymous sites than for synonymous sites in island birds compared with their mainland sister taxa (Johnson and Seger 2001), and slower rates of molecular evolution in high-elevation hummingbirds (Bleiweiss 1998). This calls for more integrated approaches in comparative molecular evolution, by including several levels of organization, different time frames, and multiple interactions among them.

One simple yet fundamental genomic feature that may benefit from such an integrated approach is DNA sequence length. Whereas the length of large genomic regions should depend largely on duplications or transpositions of DNA, the length of small DNA sequences should be conditioned mostly by the accumulation of short insertions and deletions. Plant organelle genomes (mitochondrial [mt] DNA and, especially, chloroplast [cp] DNA) are particularly well suited for such a study. Contrary to animal mtDNA, they have many intergenic spacers or introns (hereafter "spacers") that can be compared across species thanks to the conserved organization of these genomes (De Las Rivas et al. 2002). The presence of multiple copies of organelle genomes in each cell (polyploidy) and their "relaxed" transmission following cell division (Birky 1983) imply another level of selection compared to nuclear genome (Rand 2001) that might influence the evolution of organelle spacer lengths.

Here, we first ask if there have been significant shifts in organelle spacer lengths during the evolution of seed plants by testing the null hypothesis that the direction of these changes is independent across spacers in each plant lineage. Second, we examine if other features of the organelle genomes and of the plants themselves, including life history traits or level of population genetic structure, are related to spacer lengths.

Any model of sequence evolution should be compatible with population genetic principles (Lynch and Conery 2003). In particular, under the nearly neutral model of evolution (Ohta 1992), reductions in effective population size (N_e) will magnify the power of random genetic drift at the expense of selection, providing a permissive environment for the accumulation of slightly deleterious mutations. In the particular context of organelle sequences, the presence within each cell of a "population" of organelle genomes has been hypothesized to give place to a "replication race" that selects for small genomes, assuming that small genomes replicate more rapidly than larger ones (Cortopassi et al. 1992; Selosse et al. 2001). Therefore any factor that reduces effective population size of organelle genomes within cell lineages should reduce the efficacy of selection for small genomes and result in sequence growth. Conversely, any factor that increases intracellular diversity (i.e., heteroplasmy) should mitigate the effect of intracellular genetic drift and allow for the selection of more compact genomes made up of short spacers.

One factor that might increase heteroplasmy is biparental inheritance. Biparental inheritance of organelles is relatively common in plants (Harris and Ingram 1991). Moreover, the numerous shifts of plant organelles transmission inferred using a phylogenetic perspective (see Fig. 3 of Birky 1995) imply that biparental inheritance has been pervasive during plant evolution.

Large effective plant population size should also favor heteroplasmy. The rationale is that it should help maintain high levels of within-population genetic diversity. The existence of genetic differences between organelle genomes of individual plants engaged in mating is a necessary condition for biparental inheritance to result in heteroplasmy. Local effective population size is difficult to measure directly, but genetic structure could be used as a surrogate, because it typically increases when local effective population size decreases (i.e., when drift increases).

The above arguments lead to two predictions regarding the length of organelle spacers:

- Prediction 1: Spacers should be shorter in lineages that have experienced some level of biparental inheritance than in lineages where uniparental inheritance is the rule.
- Prediction 2: Spacers should be shorter in species characterized by a weak genetic structure than in species characterized by a strong genetic structure.

Competition between organelle genomes, achieved through differential replication within cell, could lead not only to changes in sequence length but also to changes in nucleotide composition. AT-rich sequences are energetically less costly than GC-rich ones (Rocha and Danchin 2002) and have been hypothesized to replicate faster (Ballard 2000). This leads to a third prediction:

- Prediction 3: Spacer lengths should be positively correlated with the GC content of organelle genomes.

All these trends should further depend on mutation rate. Owing to organelle polyploidy, mutations lead to instant heteroplasmy. Under the replication race model, heteroplasmy sets the stage for an effective selection within cell. Mutants characterized by shorter spacers should therefore increase in frequency following a few cell divisions, whereas those with longer spacers should be counterselected. In contrast, stasis will prevail in the absence of mutations. Hence the following prediction:

- Prediction 4a: Spacers should be shorter in lineages characterized by high mutation rates than in those characterized by low mutation rates.

In view of the generally negative relationship between generation time and the rate of sequence evolution (Wilson et al. 1990), we also expect the following:

- Prediction 4b: Spacers should be shorter in lineages characterized by short generation time than in those characterized by long generation time.

Finally, many of the above predictions should in principle hold for both cpDNA and mtDNA genomes of the same species, at least if their mode of transmission is identical. Hence the last prediction:

- Prediction 5: The size of cpDNA and mtDNA spacers should be positively correlated.

As plant cpDNA and mtDNA evolve slowly in sequence (Wolfe et al. 1987), broad-scale comparisons are particularly likely to reveal useful patterns. However, in any comparative study, phylogenetic relationships between the taxa sampled can be a source of statistical pseudoreplication (Freckleton et al. 2002). Therefore, to evaluate the significance of the relationships between variables, one should first test for phylogenetic independence (Abouheif 1999) and then apply comparative techniques controlling for species relatedness (Felsenstein 1985).

So far, comparative studies of molecular evolution have relied mostly on direct correlation approaches, paired sister group comparisons, or (less frequently) star phylogenies. However, the first approach means that species are inappropriately treated as independent data points, greatly increasing type I error (the risk of incorrectly rejecting the null hypothesis of no relationship among traits), particularly with large datasets (Martins and Garland 1991); the second approach amounts to analyzing only a small proportion of the existing data (Felsenstein 1985); and the third one is neither very general nor very precise, since it can be applied only to species that are approximately equally divergent. A more promising method relies instead on a set of independent contrasts distributed over the phylogenetic tree (= full tree or nested-contrasts analysis), as often performed in ecological studies (e.g., Silvertown and Dodd 1996) but rarely so in studies of molecular evolution. With this method, all possible contrasts are used ($s - 1$, compared to a maximum of $s/2$ for paired comparisons, where s is the number of species), resulting in statistically more powerful tests than conventional approaches (Ackerly 2000). Furthermore, not only recent but also deep nodes of the phylogeny are included, which can be of interest for traits that evolve slowly, while obviating the need for an arbitrary decision about the correct taxonomic level at which to make the comparisons. Finally, the assumptions made are not much more restrictive than those required by paired sister group comparisons (Martins 2000). In this study, simple and partial regression analyses based on statistically independent contrasts obtained with this method are used to test predictions 1–5.

Materials and Methods

Analysis of Spacer Lengths in Species with Completely Sequenced Chloroplast Genomes

A total of 12 vascular plants (Fig. 1) with completely sequenced chloroplast genomes were analyzed for this study: *Arabidopsis thaliana* (accession number AP00423), *Atropa belladona* (AJ316582), *Lotus corniculatus* var. *japonicus* (AP002983), *Marchantia polymorpha* (X04465),

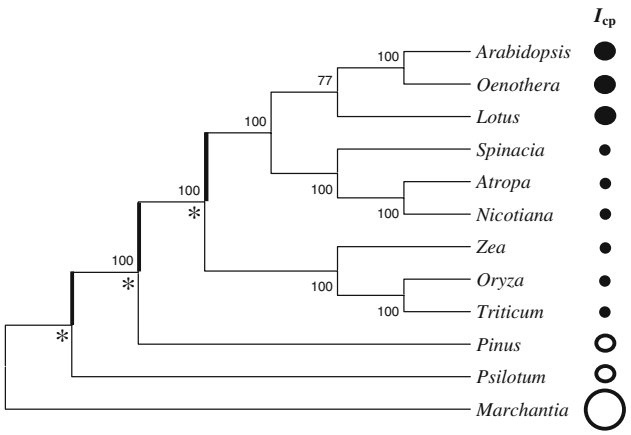

Fig. 1 Trends in spacer lengths for completely sequenced chloroplast genomes. The 50% bootstrap consensus tree was obtained from the neighbor-joining analysis of number of substitutions estimated with Kimura's (1980) two-parameter method for 56 genes combined (see Supplemental Online Material 3 for the list of genes used). Numbers on branches indicate bootstrap support from 500 replicates (in percentages). Asterisks indicate significant *t*-tests, i.e., significant directional changes in spacer lengths at a given node, with bold bars indicating the lineage with the longest spacers. Filled circles indicate I_{cp} estimates >1, and open circles estimates <1 (exact values are provided in Supplementary Online Materials 1 and 5)

Nicotiana tabacum (Z00044), *Oenothera elata* (AJ271079), *Oryza sativa* (X15901), *Pinus thunbergii* (D17510), *Psilotum nudum* (AP004638), *Spinacia oleracea* (AJ400848), *Triticum aestivum* (AB042240), and *Zea mays* (X86563). For each of these species the length of all genes, introns, and intergenic spacers, as well as their nucleotide composition, were determined using annotations of the genomes available in the organelle genome database (http://www.megasun.bch.umontreal.ca/ogmp/projects/other/cp_list.html). For plant mtDNA, this approach could not be used due to the low number of completely sequenced plant mitochondrial genomes.

Analysis of Spacer Lengths in a Broader Set of Species

In total, we selected 75 seed plant species for further studying the variation of spacer lengths (Fig. 2 and Supplemental Online Material 1). Some of these species were included because they had been investigated in population genetic surveys of cpDNA variation (e.g., 22 woody species studied by Petit et al. 2003) or because the mode of inheritance of their organelles had been described previously (Supplementary Online Material 1). The remaining species were selected to improve the taxonomic coverage. DNA was isolated from one individual in each of these 75 species using the Qiagen mini spin column extraction kit. A total of 33 consensus primer pairs amplifying noncoding DNA fragments (21 for cpDNA and 12 for mtDNA), selected among those listed by Grivet et al. (2001) and Duminil et al. (2002), were used to amplify cpDNA and

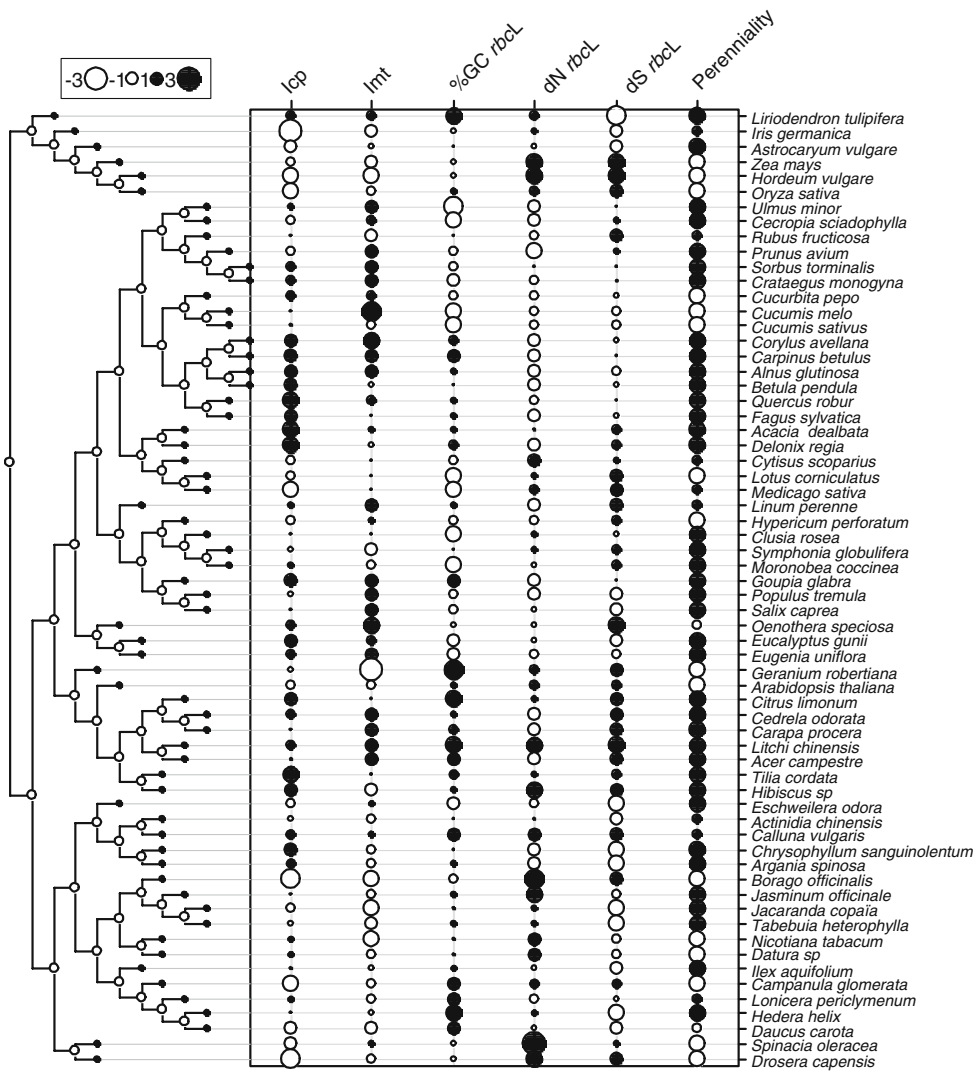

Fig. 2 Species attributes and measures of molecular evolution mapped onto the phylogeny. The size of the circles is proportional to the values of the different traits (white for low and black for high values)

mtDNA spacers (list of amplified fragments in the Supplementary Online Material 2). The PCR conditions were as described in these papers. All cpDNA fragments amplified were from the large single copy region. Amplified fragments were submitted to electrophoresis in 2% agarose gels and compared with molecular weight standards (Hartley and Donelson 1980). Fragment lengths were determined with the option Whole Band Analysis of the Bio Image software (Bio Image Systems, Inc., Jackson, MS). Three species whose organelle genomes had been completely sequenced (cpDNA, *A. thaliana* and *N. tabacum*; mtDNA, *A. thaliana* and *Beta vulgaris*) were used for comparison with gel-based estimates of the size of the amplified DNA fragments.

Index of Spacer Lengths

For the 12 species with complete chloroplast genomes available, a total of 33 conserved sequences (28 intergenic spacers and five introns) were used. They represent ~17% of the noncoding fraction of cpDNA. For the study based on direct amplification of a number of cpDNA sequences in 75 seed plant species, the fragments amplified represent, respectively, 43% and 3% of the noncoding cpDNA and mtDNA genome of *A. thaliana*. For each DNA segment in each species, the relative fragment size was determined as the ratio of the size of the fragment in this species divided by the mean size of the fragment across all species. We called I_{cp} (for cpDNA) and I_{mt} (for mtDNA) the average of these ratios across fragments. These indexes provide an indication of the relative spacer lengths for each species: $I > 1$ indicates relative obesity and $I < 1$ indicates compaction.

Phylogenetic Trees Used in Comparative Approaches

Availability of phylogenetic trees is a prerequisite for the application of comparative methods based on independent contrasts. The following trees were used. (i) For the 12 species for which a complete cpDNA sequence was available, 56

genes (listed in Supplemental Online Material 3) were separately aligned using the accessory application CLUSTALW of the BIOEDIT 5.0.9 sequence alignment editor software package (Hall 1999). The conservation of the sequences among the species was sufficient to permit largely unambiguous alignments. After removing gaps, the 56 aligned sequences were concatenated to form a single 30.12-kb sequence. A neighbor joining phylogenetic tree was obtained using PAUP 4.0b10 (Swofford 2002). *Marchantia* was used to root the tree (Fig. 1). (ii) For the 75 species studied experimentally (Fig. 2), the branching order was reconstructed from published phylogenies (Albach et al. 2001; Bruneau et al. 2001; Cuénoud et al. 2002; Gustafsson et al. 2002; Kajita et al. 2001; Soltis et al. 2000; Wojciechowski 2003). Given the various sources, it was not possible to assign branch lengths proportional to divergence time to the tree (Martins and Garland 1991). Simulations have shown that assigning the same length to all branches in comparative studies results in only a weak inflation of type I error rates (Ackerly 2000).

Testing for Shifts in Spacer Lengths During Seed Plant Evolution

Felsenstein's (1985) phylogenetically independent contrasts method was adapted to test the null hypothesis of random fluctuation of spacer lengths for both data sets (annotation of complete genomes or gel-based estimates of the size of PCR-amplified fragments). Each node on the phylogeny represents an independent comparison (11 and 74 contrasts, respectively). Starting from the tips of the tree, the hypothesis that the DNA fragments do not differ in size between the two lineages is tested using a paired *t*-test across all fragments. Relative size estimates are used to give equal weights to all fragments. The ancestral sizes of the fragments are calculated according to the procedure described by Felsenstein (1985). The same test is applied to the higher contrasts but using the reconstructed values instead. Corrections (sequential Bonferroni test [Holm 1979]) are applied to control for Type I error inflation when using multiple test procedures. For each organelle genome, the hypothesis that an overall trend exists for spacer lengths was tested by combining all *p*-values of the *t*-test applied at each contrast using a meta-analytical method (software package developed by R. Schwartzer; available at http://www.userpage.fu-berlin.de/~health/meta_e.htm). Note that it is not possible to differentiate with certainty between spacers' expansion (tendency for obesity) and spacers' reduction (tendency for compaction): the changes are only relative ones.

Independent Factors

Three species characteristics were compiled from the literature from as many of the 75 species as possible: mode of transmission of the chloroplast genome (the number of species with known mode of mitochondrial inheritance was insufficient for inclusion), fixation index estimated using cpDNA markers (G_{ST}, an estimate of F_{ST}; data were extracted from Petit et al. 2005); and perenniality, a surrogate for generation time (see Supplementary Online Material 1). The molecular variables were obtained from the analysis of two frequently sequenced chloroplast genes, *rbc*L (available in 65 of the 75 species) and *atp*B (available in 34 species). For each of these genes, we estimated GC content as well as synonymous (silent; d_S) and nonsynonymous (amino acid replacing; d_N) rates of nucleotide substitution. Substitution rates for *rbc*L and *atp*B genes were calculated with Li's (1985) method as implemented in MEGA 2.1. (Kumar et al. 2001), using *Ginkgo biloba* as reference and excluding *Pinus* (as it is sister to *Ginkgo* compared to angiosperms). *Gingko* was selected since it has a slower evolutionary rate than *Pinus* (Soltis et al. 2000) and its sequences should therefore be closer to those of the ancestor of all seed plants. We applied a test of substitution saturation (Xia et al. 2003) for both *rbc*L and *atp*B using the computer program DAMBE (Xia and Xie 2001).

Comparative Analyses

When dealing with comparative data, the first step is to test if species trait values depend on their position in the phylogeny (Blomberg et al. 2003). A useful quantitative test of the existence of phylogenetic effects is Abouheif's (1999) test for serial dependence. His mean *C*-statistic is equivalent to the statistic *I* of Moran (1948) and measures phylogenetic autocorrelation from the tree topology. Computations of phylogenetic inertia were carried out using R (Ihaka and Gentleman 1996) with the ade4 package (http://www.lib.stat.cmu.edu/R/CRAN/). G_{ST} was arcsinus square-root transformed to improve normality.

Relationships between traits were tested using the contrasts obtained with COMPARE 4.6 (Martins 2004). The five predictions (corresponding to 11 different tests) that form the core of this study (association between spacer lengths and other traits) represent different but related tests so we decided to use correction for Type I error inflation using a sequential Bonferroni procedure, as above, although this might be considered overzealous in this case (Cabin and Mitchell 2000). On the other hand, since predictions had been made regarding the sign of the relation between spacer lengths and the various traits, we used one-tailed tests for all 11 relations. For the remaining relationships, the goal was simply to identify covariates that might confound the primary relationships, so no Bonferroni corrections were used. Simple and partial regression analyses were performed with SYSTAT version 10.2 with an intercept of zero, as recommended by Pagel (1992) and using

a generalized linear model procedure. To check if the lack of information on branch lengths (all branch lengths arbitrarily set to 1) affect the results, branch lengths were transformed using Grafen and Pagel's methods that place terminal taxa at the same total height from the root (CAIC manual; available at http://www.bio.ic.ac.uk/evolve/software/caic/index.html) and the analyses were rerun. Only those relationships that remained significant after transformation were considered conclusive.

For perenniality, annuals, biannuals, short-lived perennials and long-lived perennials were coded from 1 to 4 and the character was treated as a quantitative variable. For cpDNA inheritance, we distinguished strict maternal inheritance (coded as 0) from paternal or biparental inheritance (both coded as 1), because the two genera that have paternally inherited cpDNA, *Pinus* and *Actinidia* (Supporting Online Material 4) present occasional maternal leakage (Chat et al. 2002). Moreover, Birky (1995) has suggested that paternal inheritance is a derived state that has passed by a transitory state of biparental inheritance. Hence, lineages with paternal inheritance should have experienced a period of increased frequency of heteroplasmy dating from the time when their cpDNA was biparentally inherited.

Results

Shifts in Spacer Lengths

Based on 12 completely sequenced cpDNA genomes, the number of shared noncoding fragments varies from 118 between *Nicotiana* and *Atropa* (~84% of the noncoding fragments present in each species), 72 between *Marchantia* and *Psilotum*, 124 between *Oryza* and *Triticum*, 93 between eudicots and monocots, 54 between angiosperms and gymnosperms, and 33 fragments for the highest contrast (*Marchantia* versus all other species). Three significant shifts in standardized cpDNA spacer lengths were identified, of 11 independent contrasts (see Supplementary Online Material 5): between *Marchantia* and all the other species, between *Pinus* and angiosperms, and between monocots and eudicots. The second set includes 75 species for which PCR-based estimates of spacer lengths were obtained for both cpDNA and mtDNA. A total of 21 cpDNA and 11 mtDNA fragments were amplified and sized by electrophoresis. Results for both organelles confirm and extend the previous results, as the overall tests of directional shifts in spacer lengths, based on 74 contrasts, were highly significant for both organelle genomes ($p = 6.10^{-4}$ and 5.10^{-3} for cpDNA and mtDNA sequences, respectively).

Sequence Features and Evolution

For 11 of the 12 species with completely sequenced cpDNA genome (all but *Pinus*), the proportion of noncoding sequences is lower than the proportion of coding sequences (37–49%; see Table 1). The overall GC content of the entire cpDNA genome varies from 29% (*Marchantia*) to 40% (*Oenothera*). Noncoding sequences typically have lower GC-content than coding sequences (by about 10%), but both estimates covary narrowly across species ($r = +0.92$ based on 11 independent contrasts; Supplementary Online Material 4).

Nucleotide content for *rbc*L and *atp*B can be considered to be representative of the overall cpDNA nucleotide content, as shown by investigating the 12 completely sequenced cpDNA genomes: the correlations between their GC content and the genome-wide estimates were large and significant ($r > 0.80$ for both genes; Supplementary Online Material 4).

Synonymous (d_S) and nonsynonymous (d_N) rates of nucleotide substitution based on single genes were also representative of those at other genes in the genome but

Table 1 General structural characteristics of completely sequenced chloroplast genomes

	cpDNA size (pb)	%noncoding	%GC total	%GC coding	%GC noncoding
Arabidopsis thaliana	154,478	41.1	36.7	42.2	31.1
Atropa belladona	156,687	40.9	38.2	42.7	33.6
Lotus japonicus	150,519	40.5	36.8	42.6	30.8
Marchantia polymorpha	121,024	46.5	29.1	38.9	19.0
Nicotiana tabacum	155,939	41.9	39.0	43.2	34.7
Oenothera elata	163,935	41.3	39.7	43.1	35.6
Oryza sativa	134,525	47.9	39.0	43.1	34.7
Pinus thunbergii	119,707	53.8	39.3	42.5	35.1
Psilotum nudum	138,829	36.7	35.9	41.6	29.6
Spinacia oleracea	150,725	39.5	37.2	42.2	31.6
Triticum aestivum	134,534	46.5	38.2	42.3	33.5
Zea mays	140,387	48.8	38.5	42.8	33.7

less so than GC content: the largest correlations with genome-wide estimates were $r = 0.76$ for d_N *atp*B and $r = 0.69$ for d_S *rbc*L, whereas $r < 0.50$ for d_S *atp*B and d_N *rbc*L (Supplementary Online Material 4).

Character Trends Across the Phylogeny

For several of the characters investigated, some groupings along the phylogeny were apparent: sister clades tend to have more similar values than expected by chance for most characters (Fig. 2 and Supplementary Online Material 1). For instance, Betulaceae and Fagaceae both have fairly long cpDNA spacers. The results of Abouheif's tests (1999) indicate that most traits are influenced by the position of the species in the phylogeny; only mode of cpDNA inheritance and one estimate of evolutionary rate (for *atp*B) do not present significant phylogenetic effects in this sample (Table 2).

Analyses Based on Independent Contrasts

I_{cp} and I_{mt} were then used as dependent variables in several regression analyses (Table 3). The first four predictions were supported: I_{cp} depends on mode of cpDNA transmission and on G_{ST} in the expected direction (predictions 1 and 2). I_{cp} increases with GC content at both *rbc*L and *atp*B genes (prediction 3). I_{cp} decreases when substitution rates (d_S and d_N) increase (prediction 4a), but the relationships are well supported only for *atp*B. Similarly, I_{cp} and, to a lesser extent, I_{mt} depend on perenniality (prediction 4b). On the other hand, the relationship between I_{cp} and I_{mt} (prediction 5), although positive, is not significant. As relationships between predicting variables could confound their relations with spacer lengths, additional regressions were made (Table 4). The nucleotide content at both *rbc*L and *atp*B depends neither on the substitution rate at the corresponding locus nor on perenniality (Table 4). On the contrary, all four measures of substitution rates depend on the degree of perenniality of the plant. We therefore checked whether the significant dependence of I_{cp} on substitution rate at *atp*B and *rbc*L still holds once perenniality is accounted for in partial regression analyses: this was not the case (Table 3).

Table 2 Test of phylogenetic signal for species traits

Trait	N[a]	Mean C-statistic[b]
I_{cp}	75	0.45***
I_{mt}	75	0.38***
cpDNA inheritance	42	0.06^{NS}
G_{ST}	34	0.30*
%GC *rbc*L	64	0.48***
%GC *atp*B	34	0.38***
d_S *rbc*L	64	0.42***
d_N *rbc*L	64	0.32***
d_S *atp*B	34	0.21*
d_N *atp*B	34	0.01^{NS}
Perenniality	75	0.33***

[a] Number of individuals

[b] Mean C-statistic measures phylogenetic autocorrelation; see text

NS $p < 0.95$; *$0.95 < p < 0.99$; *** p-value > 0.999

Discussion

By comparing the length of a set of noncoding spacers at each node of the two phylogenetic trees investigated, we showed that cpDNA and mtDNA spacers have experienced multiple episodes of directional changes in size (either compaction or inflation) during the radiation of plant lineages. Hence, a model of purely random variation of spacer lengths is ruled out. We are not aware of any previous studies reporting such nonrandom evolutionary trends in organelle spacer lengths.

To better understand the mechanisms underlying such directional coordinated changes, we have studied the relationships between spacer lengths and a number of plants attributes. Before carrying out these analyses, we first demonstrated that organelle spacer lengths (as measured by specifically designed indexes, I_{cp} and I_{mt}) have a strong phylogenetic inertia, although less so for mtDNA than for cpDNA. This inertia and that of several independent variables investigated justify the use of phylogenetic corrections when testing for cross-species relationships.

Following Selosse et al. (2001), we hypothesized that the existence of multiple copies of organelle genomes in each cell, which can replicate at least partly independently from each other (Birky 1983), should have important evolutionary consequences. Five predictions were tested using simple and partial regression analyses, taking advantage of all available phylogenetically independent contrasts.

Intense competition for replication between organelle genomes within cell lineages should result in compact genomes made up of short spacers. We reasoned that the intensity of this competition should be boosted by biparental inheritance, because it represents a form of gene flow between cells that results in heteroplasmy, thereby facilitating selection. We found that species with biparental cpDNA inheritance have indeed short cpDNA spacers, as predicted by the replication race model. Moreover, plant species characterized by a weak population genetic structure (low G_{ST}) also tend to have short organelle spacers. Mode of organelle transmission and G_{ST} are not independent since biparental inheritance of organelles should decrease plant

Table 3 Relationships between spacer lengths and other traits

Prediction	Dependent variable	Predictor variable	N^a	Sign	R^2	p-value[b]	Adjusted p-value[c]
1	I_{cp}	Inheritance	41	−	0.147	0.006	0.042
2	I_{cp}	G_{ST}	33	+	0.231	0.002	0.018
3	I_{cp}	%GC *rbc*L	65	+	0.131	0.001	0.015
3	I_{cp}	%GC *atp*B	35	+	0.171	0.006	0.036
4a	I_{cp}	d_S *rbc*L	63	−	0.096	0.006	$0.032^{\phi\forall}$
4a	I_{cp}	d_N *rbc*L	63	−	0.011	0.209	0.418
4a	I_{cp}	d_S *atp*B	33	−	0.188	0.005	0.040^{ϕ}
4a	I_{cp}	d_N *atp*B	33	−	0.273	0.001	0.011^{ϕ}
4b	I_{cp}	Perenniality	74	+	0.208	0.000	0.000
4b	I_{mt}	Perenniality	74	+	0.068	0.012	$0.048^{\forall}$
5	I_{cp}	I_{mt}	74	+	0.045	0.033	0.099

[a] Number of contrasts

[b] One-tailed tests

[c] Idem but p-values are adjusted using the sequential Bonferroni approach outlined in Rice (1989)

[ϕ] Relations that are no longer significant ($p > 0.05$) when perenniality is controlled for

[$\forall$] Relations that are no longer significant when using Pagel or Grafen phylogenetic tree transformation

Table 4 Relationships between predictor variables

Dependent variable	Predictor	N^a	Sign	R^2	p-value[b]
%GC *rbc*L	d_S *rbc*L	63	+	0.010	0.439
%GC *rbc*L	d_N *rbc*L	63	+	0.005	0.584
%GC *atp*B	d_S *atp*B	33	−	0.020	0.429
%GC *atp*B	d_N *atp*B	33	−	0.036	0.280
%GC *rbc*L	Perenniality	64	+	0.011	0.411
%GC *atp*B	Perenniality	35	+	0.009	0.587
d_S *rbc*L	Perenniality	63	−	0.351	0.000
d_N *rbc*L	Perenniality	63	−	0.086	$0.018^{\forall}$
d_S *atp*B	Perenniality	33	−	0.407	0.000
d_N *atp*B	Perenniality	33	−	0.323	0.000

[a] Same as in Table 3

[b] Two-tailed tests

[$\forall$] Relation that is no longer significant when using Pagel or Grafen phylogenetic tree transformation

genetic structure (Petit et al. 1993), but their combined effect on spacer lengths could not be evaluated, as these parameters were rarely available in the same species.

Species with long spacers should also be characterized by high GC content, assuming that the intracellular replication race favors not only short spacers (Selosse et al. 2001) but also high AT content (Ballard 2000) (prediction 3). In agreement with this prediction, we found that GC content at both *rbc*L and *atp*B covaried positively with spacer lengths across species. Crucially, GC content at both *rbc*L and *atp*B in species also covaried positively with the overall GC content of the cpDNA genome (see Supplementary Online Material 4; Kusumi and Tachida 2005). A key assumption here is that AT-rich genomes replicate faster than GC-rich ones, as suggested by Ballard (2000). Although nucleotide composition is known to affect basic physical aspects of the DNA molecule, such as its stability and its bendability (Vinogradov 2003), we were unable to find any published evidence supporting Ballard's suggestion of a direct link between DNA replication rate and nucleotide composition, so our hypothesis must remain tentative.

A somewhat different but related hypothesis is that of Rocha and Danchin (2002). They suggested that the higher energy costs and limited availability of G and C over A and T could explain the increased AT content of genomes that are under stringent competition for metabolic resources, such as those of symbionts or obligatory pathogens. A decreased efficacy of selection could reduce this competition for metabolic resources, resulting in the same prediction of simultaneous increase in spacer lengths and high GC content in some lineages.

Our study further uncovered a significant negative relationship between I_{cp} and substitution rates and a positive one between I_{cp} and perenniality, a surrogate for generation time (predictions 4a and 4b). Plant synonymous substitution rates have already been shown to depend on perenniality, age at maturity, or life span (Gaut et al. 1996; Laroche and Bousquet 1999; Kay et al. 2006; but see Whittle and Johnston 2003). The negative relationship between generation time and substitution rate is supported by our study. Interestingly, the relation of I_{cp} with substitution rates did not persist when perenniality was controlled for, pointing to general effects of life history on substitution rates and organelle sequence length.

The negative relation of spacer lengths with substitution rates does not depend solely on amino acid changes but holds also for synonymous substitutions, so it is unlikely to be caused by the preferential accumulation of slightly deleterious mutations in species with rapid rates of evolution. A possible interpretation of this relationship is therefore that increased rate of mutations (both substitutions and indels) should accelerate evolution toward short AT-rich sequences, through intense selection for rapid replication and/or high energetic efficiency. This implies that substitution rates and indel rates covary positively.

Recently, some evidence has accumulated showing that this could be the case. Laroche et al. (1997) have noted that for plant mtDNA, correlations are generally high between substitutions and indel events for the same intron. In rodents as well, mtDNA experiences elevated rates of substitutions and indels, compared to other mammals (Mathee et al. 2007).

The last relationship investigated was that between cpDNA and mtDNA spacer lengths (prediction 5). While positive, it was not significant. The modes of transmission of the two organelles are not always coupled in seed plants (Petit and Vendramin 2006). Furthermore, the two plant organelle genomes have evolutionary dynamics that are strikingly different from each other, especially in terms of nature and rates of mutations (Palmer 1990). As a consequence, spacer lengths of each organelle genome can vary in different directions across species.

It is remarkable that the trends we have identified for cpDNA among seed plant lineages, involving decreased GC content, decreased spacer lengths, and increased rate of evolution, parallel those observed in "resident" (obligate parasites, endosymbionts and cellular organelles) versus "free-living" genomes (Andersson and Kurland 1998; Canbäck et al. 2004). This suggests that we are dealing with an authentic syndrome and that the underlying processes are still operating long after the endosymbiotic event that gave birth to the current eukaryotic cells with their organelles. To date, the interpretation of these trends in symbionts or obligatory pathogens has not been related to the replication race model. Instead, genomic features of resident genomes have been largely viewed as maladaptive, being caused by their reduced effective population size (*Ne*), leading to increased genetic drift and the accumulation of slightly deleterious mutation at genes under less stringent selection (Woolfit and Bromham 2003). However, we note that endocellular life has typically led to polyploidy (e.g., Komaki and Ishikawa 1999), which sets the stage for intracellular competition among genome copies. Hence, the arguments put forward here to account for differences of organelle genomes among plants, based on competition within cell, might also apply to resident genomes at large.

Overall, the replication race model, whose multiple implications were first outlined by Selosse et al. (2001), although still speculative, is a good candidate to explain multiple facets of organelle genome evolution. Such a model, based on explicit processes and on general population genetic principles, appears more attractive than ad hoc evolutionary "biases." Along with an increasing body of evidence, our study also illustrates the fact that many features of organisms can drive the evolution of its genome, such as their life history and population genetic structure, justifying a broad approach in molecular evolution spanning several levels of organization, from cell to species.

Acknowledgments We are grateful to Jean Bousquet, Arndt Hampe, and Marc-André Selosse for their critical comments on an early version of the manuscript. We thank Béatrice Albert, Anne Atlan, Christian Biémont, Henri Caron, Manuela Casasoli, Deena Decker-Walters, François Delmotte, Claude dePamphilis, Laurent Duret, Joe Felsenstein, Jean-Marc Frigério, Theodore Garland Jr, Pauline Garnier-Géré, Berthold Heinze, Antoine Kremer, Emilia Martins, Brian Morton, Sophie Nadot, Carmen Palacios, David Pot, Jonathan Silvertown, and Dorothy Steane for discussions and help during this work. The research was supported by grants from the EC research program FAIR5-CT97–3795 and by the Bureau des Ressources Génétiques to R. J. Petit.

References

Abouheif E (1999) A method for testing the assumption of phylogenetic independence in comparative data. Evol Ecol Res 1:895–909

Ackerly DD (2000) Taxon sampling, correlated evolution, and independent contrasts. Evolution 54:1480–1492

Albach DC, Soltis DE, Soltis PS, Olmstead RG (2001) Phylogenetic analysis of Asterids based on sequences of four genes. Ann Mo Bot Gard 88:163–212

Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6:263–268

Ballard JWO (2000) Comparative genomics of mitochondrial DNA in *Drosophila simulans*. J Mol Evol 51:64–75

Birky CWJ (1983) Relaxed cellular controls and organelle heredity. Science 222:468–475

Birky CW Jr (1995) Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc Natl Acad Sci USA 92:11331–11338

Bleiweiss R (1998) Slow rate of molecular evolution in high-elevation hummingbirds. Proc Natl Acad Sci USA 95:612–616

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745

Bruneau A, Forest F, Herendeen PS, Klitgaard BB, Lewis GP (2001) Phylogenetic relationships in the Caesalpinioideae (Leguminosae) as inferred from chloroplast *trn*L intron sequences. Syst Bot 26:487–514

Cabin RJ, Mitchell RJ (2000) To Bonferroni or not to Bonferroni: when and how are the questions. ESA Bull 81:246–248

Canbäck B, Tamas I, Andersson SGE (2004) A phylogenomic study of endosymbiotic bacteria. Mol Biol Evol 21:1110–1122

Chat J, Decroocq S, Decroocq V, Petit RJ (2002) A case of chloroplast heteroplasmy in kiwifruit (*Actinidia deliciosa*) that is not transmitted during sexual reproduction. J Hered 93:293–300

Cortopassi GA, Shibatan D, Soong N-W, Arnheim N (1992) A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues. Proc Natl Acad Sci USA 89:7370–7374

Cuénoud P, Savolainen V, Chatrou L, Powell M, Grayer R, Chase MW (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbc*L, *atp*B, and *mat*K sequences. Am J Bot 89:132–144

De Las Rivas J, Lozano JJ, Ortiz AR (2002) Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. Genome Res 12:567–83

Duminil J, Pemonge MH, Petit RJ (2002) A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial DNA. Mol Ecol Notes 2:428–430

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712–726

Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbc*L. Proc Natl Acad Sci USA 93:10274–10279

Grivet D, Heinze B, Vendramin GG, Petit RJ (2001) Genome walking with consensus primers: application to the large single copy region of chloroplast DNA. Mol Ecol Notes 1:345–349

Gustafsson MHG, Bittrich V, Stevens PF (2002) Phylogeny of Clusiaceae based on *rbc*L sequences. Int J Plant Sci 163:1045–1054

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Harris SA, Ingram R (1991) Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plastid transmission. Taxon 40:393–412

Hartley JL, Donelson JE (1980) Nucleotide sequence of the yeast plasmid. Nature 286:860–865

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

Ihaka A, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Johnson KP, Seger J (2001) Elevated rates of nonsynonymous substitution in island birds. Mol Biol Evol 18:874–881

Kajita T, Ohashi H, Tateishi Y (2001) *rbc*L and legume phylogeny, with particular reference to Phaseoleae, Milletieae, and Allies. Syst Bot 26:515–536

Kay KM, Whittall JB, Hodges SA (2006) A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. BMC Evol Biol 6:36

Komaki K, Ishikawa H (1999) Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. J Mol Evol 48:717–722

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Kusumi J, Tachida H (2005) Compositional properties of green-plant plastid genomes. J Mol Evol 60:417–425

Laroche J, Bousquet J (1999) Evolution of the mitochondrial *rps*3 intron in perennial and annual angiosperms and homology to *nad*5 intron 1. Mol Biol Evol 16:441–452

Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Lynch M (2002) Intron evolution as a population-genetic process. Proc Natl Acad Sci USA 99:6118–23

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404

Martins EP (2000) Adaptation and the comparative method. Trends Ecol Evol 15:296–299

Martins EP (2004) COMPARE, version 4.6. Computer programs for the statistical analysis of comparative data. Department of Biology, Indiana University, Bloomington. Available at: http://www.compare.bio.indiana.edu/

Martins EP, Garland TJ (1991) Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557

Matthee CA, Eick G, Willows-Munro S, Montgelard C, Pardini AT, Robinson TJ (2007) Indel evolution of mammalian introns and the utility of non-coding nuclear markers in eutherian phylogenetics. Mol Phylogenet Evol 42:827–837

Moran PAP (1948) The interpretation of statistical maps. J Roy Stat Soc B 10:243–250

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286

Pagel MD (1992) A method for the analysis of comparative data. J Theor Biol 156:431–442

Palmer JD (1990) Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet 6:115–120

Petit RJ, Vendramin GG (2006) Plant phylogeography based on organelle genes: an introduction. In: Weiss S, Ferrand N (eds) Phylogeography of southern European refugia. Springer, New York, pp 23–97

Petit RJ, Kremer A, Wagner DB (1993) Finite island model for organelle and nuclear genes in plants. Heredity 71:630–641

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M, Mohanty A, Muller-Starck GM, Demesure-Musch B, Palme A, Martin JP, Rendell S, Vendramin GG (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300:1563–1565

Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG (2005) Comparative organisation of chloroplast, mitochondrial and nuclear diversity in plant populations. Mol Ecol 14:689–701

Rand DM (2001) The units of selection on mitochondrial DNA. Annu Rev Ecol Syst 32:415–448

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223–225

Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18:291–294

Selosse MA, Albert B, Godelle B (2001) Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol Evol 16:135–141

Silvertown J, Dodd M (1996) Comparing plants and connecting traits. In: Silvertown J, Franco M, Harper JL (eds) Plant life histories: ecology, phylogeny and evolution. Cambridge University Press, Cambridge, pp 3–16

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris JS (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbc*L, and *atp*B sequences. Bot J Linn Soc 133:381–461

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA

Vigouroux Y, Matsuoka Y, Doebley J (2003) Directional evolution for microsatellite size in maize. Mol Biol Evol 20:1480–1483

Vinogradov AE (2003) DNA helix: the importance of being GC-rich. Nucleic Acids Res 31:1838–1844

Whittle CA, Johnston MO (2003) Broad-scale analysis contradicts the theory that generation time affects molecular evolutionary rates in plants. J Mol Evol 56:223–233

Wilson M, Gaut B, Clegg M (1990) Chloroplast DNA evolves slowly in the palm family (Arecaceae). Mol Biol Evol 7:303–314

Wojciechowski MF (2003) Reconstructing the phylogeny of legumes (Leguminosae): an early 21st century perspective. In: Klitgaard BB, Bruneau A (eds) Advances in legume systematics. Part 10. Higher level systematics. Royal Botanic Garden, Kew, pp 5–35

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058

Woolfit, Bromham L (2003) Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol Biol Evol 20:1545–1555

Xia X, Xie Z (2001) DAMBE: data analysis in molecular biology and evolution. J Hered 92:371–373

Xia XH, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylog Evol 26:1–7