Coevolution between Nuclear-Encoded DNA Replication, Recombination, and Repair Genes and Plastid Genome Complexity

Jin Zhang[1], Tracey A. Ruhlman[1], Jamal S. M. Sabir[2], John Chris Blazier[1], Mao-Lun Weng[1], Seongjun Park[1], and Robert K. Jansen[1,2,*]

[1]Department of Integrative Biology, University of Texas at Austin

[2]The Biotechnology Research Group, Department of Biological Sciences, Faculty of Science, King Abdulaziz University (KAU), Jeddah, Saudi Arabia

*Corresponding author: E-mail: jansen@austin.utexas.edu.

Accepted: February 13, 2016

Data deposition: Plastid genome sequences of *Monsonia marlothii* and *Monsonia emarginata* have been deposited at NCBI under the accessions KT692738–9. SRA accession numbers for the transcriptome data for Geraniales were reported in Ruhlman et al. (2015).

Abstract

Disruption of DNA replication, recombination, and repair (DNA-RRR) systems has been hypothesized to cause highly elevated nucleotide substitution rates and genome rearrangements in the plastids of angiosperms, but this theory remains untested. To investigate nuclear–plastid genome (plastome) coevolution in Geraniaceae, four different measures of plastome complexity (rearrangements, repeats, nucleotide insertions/deletions, and substitution rates) were evaluated along with substitution rates of 12 nuclear-encoded, plastid-targeted DNA-RRR genes from 27 Geraniales species. Significant correlations were detected for nonsynonymous (d*N*) but not synonymous (d*S*) substitution rates for three DNA-RRR genes (*uvrB/C*, *why1*, and *gyrA*) supporting a role for these genes in accelerated plastid genome evolution in Geraniaceae. Furthermore, correlation between d*N* of *uvrB/C* and plastome complexity suggests the presence of nucleotide excision repair system in plastids. Significant correlations were also detected between plastome complexity and 13 of the 90 nuclear-encoded organelle-targeted genes investigated. Comparisons revealed significant acceleration of d*N* in plastid-targeted genes of Geraniales relative to Brassicales suggesting this correlation may be an artifact of elevated rates in this gene set in Geraniaceae. Correlation between d*N* of plastid-targeted DNA-RRR genes and plastome complexity supports the hypothesis that the aberrant patterns in angiosperm plastome evolution could be caused by dysfunction in DNA-RRR systems.

Key words: plastid genome rearrangement, nucleotide insertions/deletions, substitution rates, nuclear–plastid genome coevolution.

Introduction

Plastid genomes (plastomes) of angiosperms are generally highly conserved with a quadripartite structure that includes large single copy (LSC) and small single copy regions and an inverted repeat (IR) with conserved gene content and order (Ruhlman and Jansen 2014). However, several unrelated lineages have experienced substantial variation in genome organization, including Campanulaceae (Cosner et al. 2004; Haberle et al. 2008; Knox 2014), Ericaceae (Fajardo et al. 2012; Martínez-Alberola et al. 2013), Geraniaceae (Chumley et al. 2006; Guisinger et al. 2008, 2011; Blazier et al. 2011; Weng et al. 2014), and Fabaceae (Milligan et al. 1989; Perry et al. 2002; Cai et al. 2008; Sabir et al. 2014; Schwarz et al. 2015). In addition, correlation between the degree of plastome rearrangement and rates of nucleotide substitution has been suggested for several lineages (Jansen et al. 2007; Guisinger et al. 2008; Sloan et al. 2012; Weng et al. 2014). The cause of the correlation is not clear but alteration in DNA repair and recombination mechanisms has been hypothesized.

In Geraniaceae, variation in plastome complexity is unprecedented. A large number of genome rearrangements have been detected in *Hypseocharis*, *Pelargonium*, *Monsonia*,

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Geranium, and *Erodium*, including IR loss in *Monsonia* and *Erodium* and an order of magnitude in IR size difference (7–75 kb) within the family (Chumley 2006; Blazier et al. 2011; Guisinger et al. 2011; Weng et al. 2014). Plastid-encoded genes display highly elevated nonsynonymous substitution rates (d*N*) in Geraniaceae (Guisinger et al. 2008), and repetitive DNA is prevalent in the rearranged genomes (Chumley et al. 2006; Guisinger et al. 2011; Weng et al. 2014). Although dysfunction of DNA replication, recombination and repair (DNA-RRR) systems has been suggested to cause these unusual plastome phenotypes, this has not been tested (Jansen et al. 2007; Guisinger et al. 2008; Weng et al. 2014).

Angiosperm plastomes do not encode any DNA-RRR proteins to maintain genome stability (Bock 2007; Jansen and Ruhlman 2012). These genes reside in the nuclear genome and their products are imported into plastids. A number of DNA-RRR genes have been verified experimentally (Day and Madesis 2007; Maréchal and Brisson 2010; Boesch et al. 2011). For example DNA Gyrase, involved in the relaxation of DNA supercoiling, is critical for plastid DNA replication (Wang 1996; Cho et al. 2004). Two genes (*gyrA* and *gyrB*) encoding subunits of Gyrase are known in angiosperms, and plastid-targeted genes for both types are present in *Arabidopsis* and *Nicotiana benthamiana* (Cho et al. 2004; Wall et al. 2004). In DNA recombination, loss of *chloroplast mutator* (*msh1*), a homolog of the bacterial gene involved in DNA recombination and mismatch repair, has been shown to affect plastid DNA recombination and/or repair possibly by influencing the frequency or outcome of double strand breaks (Xu et al. 2011). Knockout studies in *Arabidopsis* also suggested that Whirly proteins are involved in maintaining plastome stability by suppressing recombination between short repeats, and two proteins from this family (*why1* and *why3*) are targeted to plastids (Maréchal et al. 2009). Solved structures of mitochondrial Why2 in complex with single-stranded DNA suggest that Whirly proteins suppress error-prone microhomology-mediated recombination (MHMR) via nonspecific binding to single-stranded DNA (Cappadocia et al. 2010, 2013). Surprisingly, in the organelle genomes of wild-type *Arabidopsis*, short range rearrangements (<1,000 bp) occur randomly and more frequently than previously supposed. Furthermore, MHMR and microhomology-independent repair products accumulate to similar levels in the organelle genomes of wild-type plants. The shift in repair products observed in the plastomes of Whirly knockout lines suggests that these proteins mainly suppress rearrangements arising from illegitimate recombination between proximal microhomologous regions but have little effect on homology-independent repair (Zampini et al. 2015).

Previous authors (Jansen et al. 2007; Guisinger et al. 2008, 2011; Weng et al. 2014) have hypothesized that the highly rearranged plastomes and accelerated d*N* in Geraniaceae may have resulted from aberrant DNA-RRR mechanisms. In this study, plastome complexity was estimated for 27 species in Geraniales using several independent metrics. Nuclear transcriptomes for the same species (Ruhlman et al. 2015) were mined to extract nuclear-encoded, plastid-targeted DNA-RRR genes along with a control data set comprising genes with different subcellular locations (plastid, mitochondrial, and other). Correlation between nucleotide substitution rates of three DNA-RRR genes (*uvrB/C*, *why1*, and *gyrA*) and plastid genome complexity was detected.

Materials and Methods

DNA Isolation, Genome Sequencing, and Assembly

Total genomic DNA of *Monsonia emarginata* and *Monsonia marlothii* was isolated from emergent leaves and was sequenced with Illumina HiSeq 2000 at the University of Texas Genomic Sequencing and Analysis Facility (UT GSAF) as described in Weng et al. (2014). Approximately 60 and 23 million 100-bp paired-end reads were generated from 800-bp insert libraries for *M. emarginata* and *M. marlothii*, respectively. A 10-kb SMRT cell library was constructed for PacBio RS II sequencing and one cell of sequence data was generated for each species at the University of Florida Interdisciplinary Center for Biotechnology Research. All PacBio reads were corrected with the long read error correction tool (LSC) (http://www.healthcare.uiowa.edu/labs/au/LSC/LSC_manual.html, last accessed February 26, 2016) using Illumina paired-end reads.

De novo assembly of the Illumina paired-end reads was performed with Velvet v 1.2.07 (Zerbino and Birney 2008) with various parameters (kmer 79–95 bp and coverage 200, 500, and 1,000) to optimize contig length. Corrected PacBio reads (supplementary table S1, Supplementary Material online) were used to join Velvet assemblies and create scaffolds. Both Illumina and PacBio reads were mapped back to the scaffolds using Geneious (Biomatters; http://www.geneious.com/) to fill gaps.

RNA Isolation and Transcriptome Sequencing and Assembly

Total RNA was extracted from emergent leaves of 27 species of Geraniales, and four different tissues (emergent and expanded leaves, roots, and flowers) of *Pelargonium* × *hortorum* as described in Zhang et al. (2013). Transcriptome sequencing was performed with HiSeq 2000 at UT GSAF. Sequence data were preprocessed and assembled as described in Zhang et al. (2013, 2015).

Plastid Genome Complexity Analysis

Plastid genomes of 27 species of Geraniales and *Arabidopsis* were used for the genome complexity analysis, 26 of which were downloaded from National Center for Biotechnology Information (NCBI) (*Arabidopsis thaliana* NC_000932, *Francoa sonchifolia* NC_021101, *Melianthus villosus* NC_023256, *Hypseocharis bilobata* NC_023260, *Pelargonium nanum*

KM527896, *Pelargonium citronellum* KM527888, *Pelargonium echinatum* KM527891, *Pelargonium incrassatum* KM527894, *Pelargonium fulgidum* KM527893, *Pelargonium cotyledonis* KM459516, *Pelargonium australe* KM459517, *Pelargonium dichondrifolium* KM459515, *Pelargonium exstipulatum* KM527892, *Pelargonium myrrhifolium* KM527895, *Pelargonium tetragonum* KM527899, *Pelargonium transvaalense* KM527900, *P.* × *hortorum* NC_008454, *Geranium maderense* KT760576, *Geranium phaeum* KT760577, *Geranium incanum* KT760575, *California macrophylla* JQ031013, *Erodium texanum* NC_014569, *Erodium chrysanthum* NC_027065, *Erodium gruinum* NC_025907, *Erodium foetidum* KF771022, and *Erodium trifolium* NC_024635). Plastid genomes of the remaining two species (*M. emarginata* KT692738, *M. marlothii* KT692739) were assembled in this study. Different measures of genome complexity (genome rearrangements, repeat content, insertions and deletions (indels), and substitution rates, described below) were estimated by comparing the plastid genomes of 27 species of Geraniales to *Arabidopsis*.

Multiple genome alignment of the 27 species of Geraniales and *Arabidopsis*, and pairwise genome alignment between each species of Geraniales and *Arabidopsis* were performed using the progressive Mauve algorithm (Darling et al. 2010) in Geneious. Genes shared across the 28 species were identified by a custom Python script. The locally collinear blocks (LCBs) identified by Mauve alignment and the order of the shared genes identified by a custom Python script were numbered for genome rearrangement estimation. Two genome rearrangement measures, inversion (IV) and breakpoint (BP) distances, were estimated by comparing the numbered LCBs and gene order between 27 species of Geraniales with *Arabidopsis*, using Grimm (Tesler 2002a, 2002b) and the online web server Common Interval Rearrangement Explorer (Bernt et al. 2005), respectively.

Each plastid genome was blasted against itself with NCBI-BLAST (Basic Local Alignment Search Tool) (BLAST 2.2.28+) using default parameters. One IR was removed from the plastid genomes where two copies were present. BLAST results were parsed with a custom Python script to identify dispersed repeats (DR). Tandem repeats were identified using Tandem Repeat Finder v 4.07b, (Benson 1999) with default parameters. Variation in repeat content was estimated by subtracting the number of repeats in *Arabidopsis* from the number identified in the other 27 species.

Shared protein-coding genes, intron regions, and intergenic (IG) regions among the 28 species were identified and sequences were aligned with MAFFT (Katoh and Standley 2013) in Geneious. Indels within these regions were calculated by comparing the aligned regions of 27 species in Geraniales to *Arabidopsis* using a custom Python script. For protein-coding genes, only intact genes were considered (in-frame indels). Alignments of the shared protein-coding genes were concatenated for plastid genome rate estimation.

Evolutionary Rate Estimation

PAML's codeml (Yang 2007) was used to estimate synonymous (d*S*) and nonsynonymous (d*N*) substitution rates using the codon frequencies model F3X4. Gapped regions were excluded with parameter "cleandata = 1." Pairwise rates were estimated with parameter "runmode = -2." Out of the 70 *Arabidopsis* DNA-RRR genes, 34 were identified with either experimental or computational evidence of plastid targeting and were used as reference sequences (supplementary data file S1, Supplementary Material online). To investigate presence or absence, reference genes were used for reciprocal BLAST against the transcriptomes of all Geraniales species with a cutoff of *e*-value 1e-10 as described in Zhang et al. (2013). To be counted as present in Geraniales, transcripts were required to have either a minimum length of 1 kb or represent greater than 60% of the total length of reference gene. If a reference sequence homolog was identified in all Geraniales species, it was used for evolutionary rate estimation. Ultimately 12 nuclear-encoded, plastid-targeted DNA-RRR genes were analyzed. Ninety nuclear control genes with different subcellular locations (plastid 30, mitochondrial 30, other 30) were selected from the APVO database (Duarte et al. 2010), extracted from Geraniales transcriptomes with reciprocal BLAST as described in Zhang et al. (2013) and used as negative control groups. Genes whose products interact directly with plastid encoded proteins were excluded. Fifty-nine plastid-encoded genes were extracted from the plastid genomes as described in Weng et al. (2014). The accession numbers and descriptions of corresponding genes in *A. thaliana* are shown in supplementary data file S1, Supplementary Material online.

Analysis of Correlation between Evolutionary Rate and Genome Complexity

Correlation between d*N* or d*S* of each gene and the measures of genome complexity was performed using the original mirror tree method as described in Pazos et al. (1997, 2005) and Pazos and Valencia (2001). The evolutionary rates (d*N* and d*S*) of each gene, and each genome complexity measure for the 27 species of Geraniales were collected as a rate or genome complexity vector, respectively. Correlation between the rate vector and genome complexity vector was estimated with the Pearson correlation test using built in function pearsonr in scipy module of Python. The resulting *P* values were Bonferroni corrected to remove the effect of multihypothesis testing.

Rate Comparisons between Geraniaceae and Brassicaceae

Based on previously published phylogenies of Brassicaceae (Kagale et al. 2014), nine species of Brassicaceae (*Arabidopsis lyrata*, *A. thaliana*, *Arabis alpina*, *Barbarea verna*, *Brassica napus*, *Brassica rapa*, *Capsella bursa-pastoris*,

Pachycladon cheesemanii, and *Raphanus sativus*) and one outgroup species from the Brassicales (Caricaceae *Carica papaya*) were selected. For all species included, either complete or partial plastid genome and transcriptome information was available. Reciprocal BLAST searches were used to identify orthologs of DNA-RRR and nuclear control genes. Rates for those genes present in both Geraniaceae and Brassicaceae data sets were compared. To avoid bias in the rate comparisons, a similar number (11) of species of Geraniales representing the major lineages (*Me. villosus*, *Hypseocharis bilobata*, *Pelargonium nanum*, *P. exstipulatum*, *P. myrrhifolium*, *P.* × *hortorum*, *M. emarginata*, *G. maderense*, *Erodium chrysanthum*, and *E. foetidum*) were selected. Genes were parsed into five groups based on their functions or subcellular locations (DNA-RRR, nuclear-encoded plastid-targeted control, nuclear-encoded mitochondrial-targeted control, other nuclear-encoded control, and plastid encoded). Rates of the same gene groups from the two families were compared using Student's *t*-test and Wilcoxon rank-sum test.

Results

Plastid Genome Sequencing

Illumina paired-end and PacBio reads were generated and assembled for two species of *Monsonia* (see supplementary table S1 for read information, Supplementary Material online). The IR was absent from the plastomes of *M. emarginata* (156,877 bp) and *M. marlothii* (134,416 bp). Both genomes contained 107 genes, including 74 protein-coding, 29 transfer RNA and 4 ribosomal RNA (rRNA) genes, and introns were detected in 14 genes. The GC content for *M. emarginata* and *M. marlothii* was 40.2% and 39.3%, respectively.

Genome Complexity

Four categories of genome complexity (genome rearrangements, repeat content, nucleotide substitution rates, and indels) were estimated by comparing the plastomes of 27 Geraniales species with the reference genome *Arabidopsis* (table 1 and supplementary fig. S1, Supplementary Material online), which has the same gene order as the ancestral genome of Geraniales (Weng et al. 2014). For estimating genome rearrangements, LCBs were identified by either Mauve multiple genome alignments of 27 species of Geraniales and *Arabidopsis* or by pairwise genome alignment between each of the Geraniales species and *Arabidopsis*. An additional measure of genome rearrangement was based on synteny of 63 shared genes across the 27 species of Geraniales and *Arabidopsis*.

Two metrics of genome rearrangement, BP and IV distances, were estimated based on the order of LCBs or the synteny of shared genes (supplementary fig. S2, Supplementary Material online). Within the same species, the estimated BP and IV distances were similar. The BP and IV distances of Geraniaceae species ranged from 2 to 18, whereas the outgroup species (*F. sonchifolia*, *Me. villosus*), which have similar plastid genome organization to *Arabidopsis*, had much smaller BP and IV distances (0–3). The greatest IV distance was 17 between *G. maderense* and *Arabidopsis*, and the largest BP distance was 18 between both *G. maderense* and *E. chrysanthum* and *Arabidopsis*. Multiple clade-specific increases of BP and IV distances were observed in Geraniaceae (*Hypseocharis*, *Geranium*, *Monsonia*, *Pelargonium* C2 clade and *Erodium* clade I).

Two classes of repeats, DRs and tandem repeats, were identified (supplementary fig. S3*A*, Supplementary Material online). Across all taxa DRs ranged in size from 15 to 4,377 bp with an average repeat length of 100 bp. Tandem repeats ranged in length from 25 to 3,217 bp and an average size of 100 bp. The greatest numbers of DRs (752) and tandem repeats (107) were found in *G. incanum*. The fewest repeats were found in the outgroup species for both DR (*F. sonchifolia*, 36; *Me. villosus*, 44) and tandem repeats (*F. sonchifolia*, 19; *Me. villosus*, 15). Clade-specific increases in repeat content were observed in Geraniaceae (*Geranium*, *Monsonia*, *Pelargonium* C2 clade and *Erodium* clade I).

The number of indels was estimated for protein coding and rRNA (CDR), intron and IG regions (supplementary fig. S3*B* and table S2, Supplementary Material online). Among the 27 species of Geraniales and *Arabidopsis*, 63 CDRs, 9 intron and 24 IG regions were identified as shared. The greatest number of IG indels (364) was identified in *G. phaeum* and *G. incanum*. CDR indels (123) were most abundant in *M. emarginata*, and *E. chrysanthum* had the most intron indels (185). Fewer CDR indels (33–37) were identified in the outgroup species compared with Geraniaceae species (73–123) and the number of intron and IG indels was similar across all species. The size of CDR indels ranged from 3 to 411 bp (average 13 bp). For IG indels ranged in size from 1 to 481 bp (average 10 bp) whereas indels of introns ranged from 1 to 116 bp (average 6 bp).

Nucleotide Substitution Rates

Plastome synonymous (d*S*) and nonsynonymous (d*N*) substitution rates were estimated using the concatenated alignment of 59 shared protein genes (supplementary table S2, Supplementary Material online; CDR except those in bold). The outgroup species had lower d*N* (0.039–0.041) and d*S* (0.35–0.37) relative to Geraniaceae (d*N*, 0.065–0.094; d*S*, 0.43–0.62). The highest d*N* (0.094) was in *E. chrysanthum*, whereas *E. chrysanthum* and *E. gruinum* had the highest d*S* (0.62; table 1). Multiple clade-specific accelerations of d*N* were detected within the *Pelargonium* C clade and in *Erodium* clade I (supplementary fig. S1, Supplementary Material online). Clade-specific acceleration of d*S* was only observed within *Erodium* clade I, whereas other Geraniaceae species had a similar d*S* value (~0.5).

Table 1

Measures of Genome Complexity among 27 Geraniales Species

Species	LCBs-p	Gene Order	DR	Tandem	CDR	Intron	IG	d*N* CP	d*S* CP
F. sonchifolia	0	0	36	19	33	132	283	0.039	0.35
Melianthus villosus	2	0	44	15	37	150	299	0.041	0.37
H. bilobata	12	11	88	55	73	158	292	0.065	0.43
P. nanum	5	3	99	47	109	152	331	0.086	0.50
P. citronellum	5	3	84	40	108	153	336	0.087	0.50
P. echinatum	5	3	71	33	110	155	326	0.088	0.50
P. incrassatum	5	3	125	56	106	152	339	0.088	0.50
P. fulgidum	5	3	116	43	108	151	340	0.087	0.50
P. cotyledonis	6	4	87	28	106	156	337	0.087	0.50
P. australe	6	4	90	42	105	155	338	0.086	0.50
P. dichondrifolium	7	4	105	42	104	154	330	0.087	0.50
P. exstipulatum	6	4	120	47	106	159	342	0.087	0.50
P. myrrhifolium	8	4	79	31	115	154	332	0.091	0.50
P. tetragonum	7	3	72	27	115	158	329	0.090	0.50
P. transvaalense	12	8	490	41	107	156	324	0.087	0.50
P. x *hortorum*	11	12	171	25	120	156	325	0.093	0.51
M. emarginata	9	13	191	58	123	169	333	0.088	0.52
M. marlothii	12	14	160	85	117	173	336	0.085	0.50
G. maderense	13	17	196	38	114	171	310	0.081	0.52
G. phaeum	11	10	378	64	120	174	364	0.079	0.51
G. incanum	14	11	752	107	122	175	364	0.078	0.50
C. macrophylla	3	4	58	40	79	165	311	0.071	0.49
E. texanum	12	12	132	27	110	180	300	0.083	0.52
E. chrysanthum	15	11	122	33	110	185	326	0.094	0.62
E. gruinum	14	9	72	35	111	166	319	0.093	0.62
E. foetidum	4	3	58	46	88	172	306	0.076	0.50
E. trifolium	6	5	60	35	95	170	320	0.077	0.51

NOTE.—Measures of genome complexity that were highly correlated to other complexity measures are not shown (supplementary tables S3 and S4, Supplementary Material online). F, *Francoa*; H, *Hypseocharis*; P, *Pelargonium*; M, *Monsonia*; G, *Geranium*; C, *California*; E, *Erodium*. LCBs-p, inversion distance estimated from local collinear blocks; Gene order, inversion distance estimated from gene order; DR, small dispersal repeats; Tandem, repeats estimated from Tandem Repeat Finder (Benson 1999); CDR, indels in coding and rRNA regions; IG, indels in intergenic regions; Intron, indels in intron regions.

Of the 70 DNA-RRR genes evaluated, 34 were predicted to encode a plastid targeting peptide. Among those, 12 were present in all sampled species (see Materials and Methods) (supplementary data file S1, Supplementary Material online). Here, the Geraniales homologs with putative functional assignments in *Arabidopsis* have been referred to according to their conserved domains. The value of d*S* and d*N* for each of the 12 nuclear-encoded DNA-RRR genes, 90 nuclear-encoded control genes, and 59 plastid-encoded genes was estimated based on the MAFFT alignments (supplementary data file S2, Supplementary Material online). The 90 control genes were divided into three groups based on their subcellular localization (plastid 30, mitochondrial 30, other 30; supplementary data file S1, Supplementary Material online). Of the five gene groups, DNA-RRR genes, plastid-encoded genes, nuclear-encoded plastid-targeted genes (NUCP), nuclear-encoded mitochondrial-targeted genes (NUMT) and other nuclear-encoded genes (NUOT), NUMT had the highest median value of d*N* (0.19) and d*S* (2.14; fig. 1). Among nuclear-encoded genes, NUCP had the lowest d*N* (0.17), whereas NUOT had the lowest d*S* values (2.00). Both d*N* (0.061) and d*S* (0.52) of plastid-encoded genes were much lower than nuclear genes (fig. 1). The average d*N* of nuclear genes (0.18) was approximately three times higher than plastid-encoded genes (0.061), and d*S* was about four times higher (2.09 vs. 0.52).

Correlation of Genome Complexity and Nucleotide Substitution Rates

To evaluate the correlation between measures of plastid genome complexity and evolutionary rates in DNA-RRR genes, each metric was calculated for 27 species as a vector. First, tests among plastome complexity measures were performed to eliminate those that were highly correlated with each other. Of the six measures of genome rearrangement (supplementary table S3, Supplementary Material online) four were highly correlated (correlation coefficient >0.95). This resulted in the elimination of all but two genome rearrangement metrics, IV distance based on either LCBs from pairwise genome alignment (LCB-p) or on synteny of shared genes (gene order). Correlations among all remaining measures of genome complexity were calculated

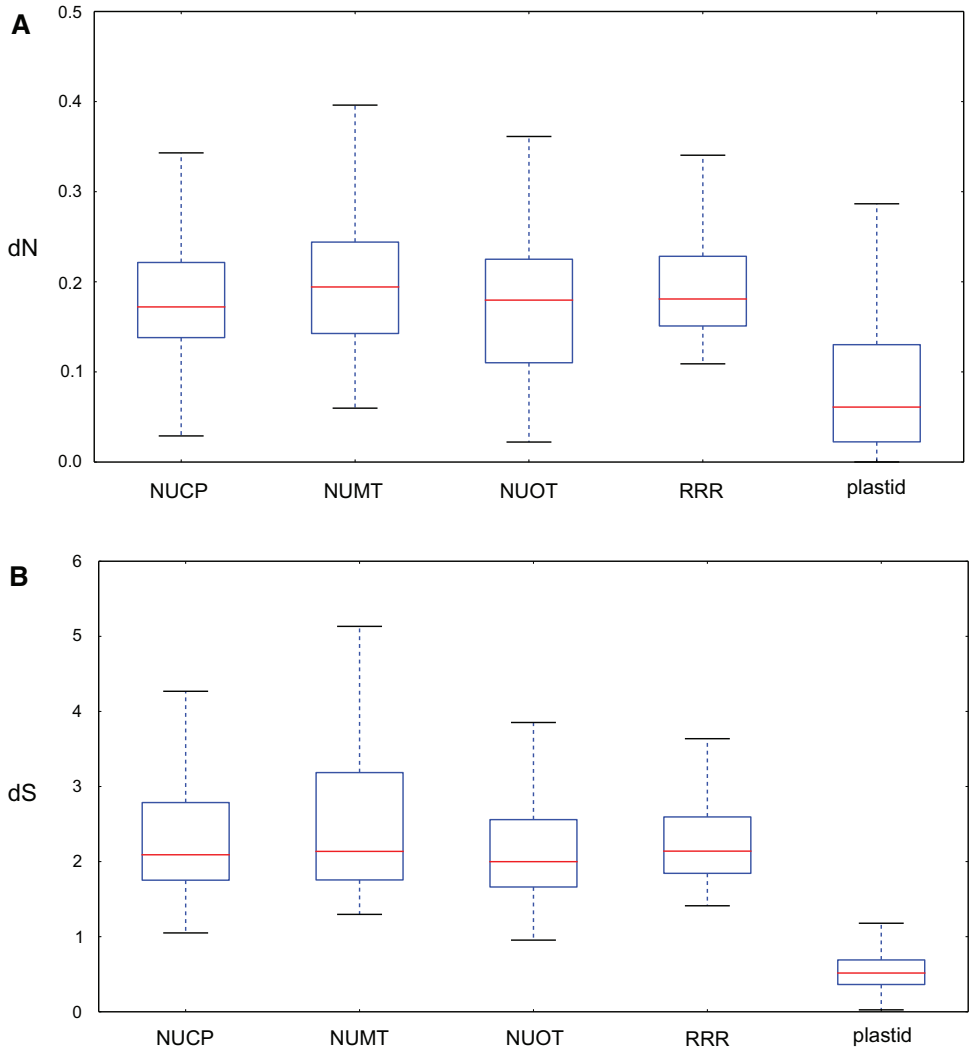

Fig. 1.—Evolutionary rates of DNA-RRR, nuclear control, and plastid-encoded genes. (*A*) Nonsynonymous (d*N*) and (*B*) synonymous (d*S*) substitution rates of different gene groups are shown. NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear encoded control genes; RRR, plastid-targeted DNA replication, recombination and repair genes; plastid, plastid-encoded genes. Boxes represent sorted rates ranging from 25% to 75%, dashed lines extend to maximal and minimal rates, red lines represent the median rate for each gene group.

(supplementary table S4, Supplementary Material online). The highest correlation coefficient was observed between CDR indels and d*N* (0.92). High correlation coefficient was also observed between LCB-p and gene order (0.88), and between plastome d*N* (d*N* CP [nonsynonymous substitution rates of the plastid genome]) and d*S* (0.83). The correlation coefficients among the remaining measures ranged from −0.24 to 0.77.

Correlation between each of the nine remaining measures of plastome complexity and evolutionary rates of the 12 nuclear-encoded DNA-RRR genes, along with the control data set, was evaluated for significance (Pearson correlation test, $P < 0.05$ after multihypothesis correction; fig. 2). Unless otherwise stated only correlations found significant are reported in the text. Correlation of d*S* and genome complexity was detected in two pairs: Between *recB* and Intron indels, and between *odb2* and DR (fig. 2*B*). The number of genes with d*N* correlated to genome complexity measures is summarized in figure 3. Details on all pairs showing significant correlation of d*N* are in supplementary table S5, and results for all comparisons are in supplementary data file S3, Supplementary Material online.

Three genes were identified with correlation to genome complexity in each of the NUMT and DNA-RRR gene groups, and 10 of the 30 NUCP genes were correlated with genome complexity (fig. 3). Among the correlated DNA-RRR genes was a homolog encoding the uvrB/uvrC motif (AT2G03390). This gene in Geraniales will be referred to hereafter as *uvrB/C*. Among the DNA-RRR genes, *uvrB/C* was

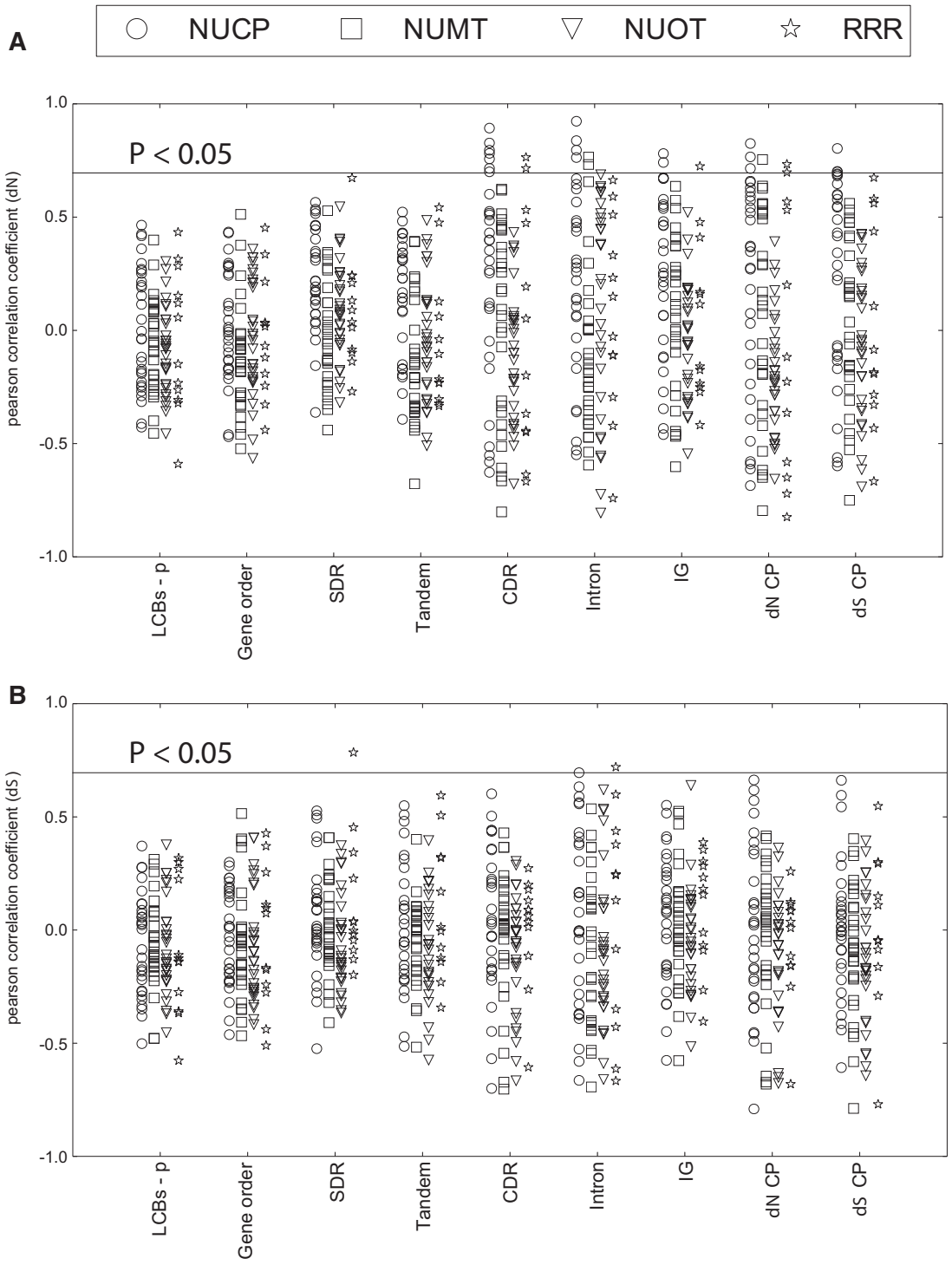

Fig. 2.—Pearson correlation coefficient between gene evolutionary rates and genome complexity. Correlations of (*A*) nonsynonymous (d*N*) and (*B*) synonymous (d*S*) substitution rates of genes and genome complexity are shown. NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear encoded control genes; RRR, plastid-targeted DNA replication, recombination and repair genes; LCBs-p, inversion distance estimated from local collinear blocks; Gene order, inversion distance estimated from gene order; SDR, small dispersal repeats; Tandem, repeats estimated from Tandem Repeat Finder (Benson 1999); CDR, indels in coding and rRNA regions; Intron, indels in intron regions; IG, indels in intergenic regions; d*N* CP, nonsynonymous substitution rates of the plastid genome; d*S* CP, synonymous substitution rates of the plastid genome.

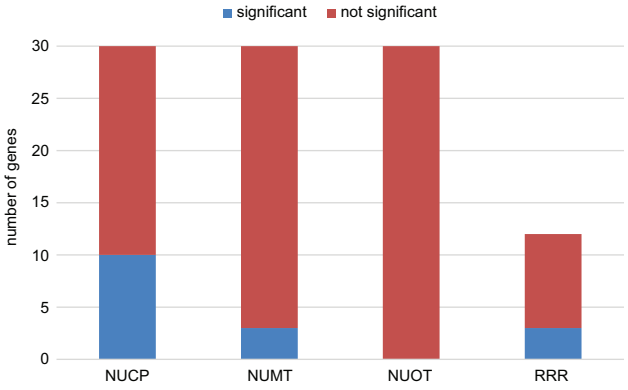

Fig. 3.—Significant correlation of evolutionary rates (d*N*) and genome complexity in different gene groups. Significant correlations were identified in NUCP, NUMT, and RRR but not NUOT genes. NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear encoded control genes; RRR, plastid-targeted DNA replication, recombination and repair genes.

correlated with CDR and IG indels, and d*N* CP. Correlation was also detected between *gyrA* and d*N* CP, and between *why1* and CDR indels (supplementary table S5, Supplementary Material online). No significant correlation was identified using measures of genome rearrangement or repeat content. Of the measures of genome complexity showing correlation, the greatest number (9) was for CDR indels. Although three NUMT genes showed a correlation with various measures of plastid genome complexity, two are targeted to both plastids and mitochondria (supplementary table S5 and data file S3, Supplementary Material online).

To further investigate the differences between NUCP and DNA-RRR genes, d*N* of the ten NUCP and three DNA-RRR genes showing significant correlation to genome complexity was compared with the remaining 20 NUCP and 9 DNA-RRR genes, respectively. The average d*N* was higher in the correlated gene sets (fig. 4), but the difference was only significant for NUCP genes (Wilcoxon rank-sum test, $P < 0.05$). No significant difference of d*S* was identified.

Rate Acceleration in Geraniales

Significant correlations between d*N* and genome complexity in NUCP control genes were unexpected; therefore, another angiosperm order was investigated to compare nucleotide substitution rates in plastid- and nuclear-encoded genes. Similar gene sets from ten species of Brassicales were assembled from published data (supplementary table S6, Supplementary Material online). Five DNA-RRR, 28 NUCP, 19 NUMT, 20 NUOT, and 59 plastid-encoded genes common to both Geraniales and Brassicales were identified (supplementary table S7, Supplementary Material online). Because the Geraniales data set comprised a greater number of taxa, a subset was utilized for rate comparisons (see Materials and Methods). Acceleration of d*N* was observed in all gene groups in Geraniales compared with Brassicales (fig. 5*A*), and significant acceleration (Student *t*-test, $P < 0.05$) was observed in NUCP, RRR, and plastid-encoded genes. Significant acceleration of d*S* in Geraniales compared with Brassicales was identified in plastid-encoded genes but not in any nuclear-encoded genes (fig. 5*B*).

Discussion

Nucleotide Substitution Rates in Geraniaceae

Synonymous (d*S*) and nonsynonymous (d*N*) substitution rates of the nuclear- and plastid-encoded genes were estimated in Geraniaceae. The ratio of d*N* and d*S* in nuclear to plastid protein-coding genes was approximately 3:1 and 4:1, respectively (fig. 1). These values fall within the range of previous estimates for d*N* (1:1 to 5:1) and d*S* (4:1 to 6:1) (Wolfe et al. 1987; Gaut 1998; Drouin et al. 2008) and demonstrate that, across angiosperms, the ratio of d*N* between the nuclear and plastid genomes fluctuates more than that of d*S*. This variation may reflect lineage-specific effects, which are more likely to affect d*N* (Wolfe et al. 1987; Wolf 2012).

Rates of nucleotide substitutions in nuclear genes with products targeted to different subcellular locations, plastid (NUCP), mitochondrion (NUMT) and other (NUOT), were very similar (d*N* 1.0:1.0:1.1, d*S* 1.1:1.1:1.0; fig. 1) in Geraniaceae. One attribute of these analyses was the family-wide nature of the sampling allowing the phylogenetic context necessary to assess lineage-specific effects. Sloan et al. (2014) compared rates of ribosomal proteins with different subcellular locations (plastid, mitochondria, cytosol) using two pairs of *Silene* species. Although results were similar in d*S* (1.03:1.03:1), unlike Geraniaceae, the d*N* values of cytosol-targeted ribosomal proteins were much lower in species with fast evolving organelle genomes than their organelle-targeted counterparts in *Silene* (14:12:1). It is possible, as NUOT genes could be targeted anywhere except the plastid and mitochondrion (e.g., endoplasmic reticulum or nucleus), that a much higher d*N* in the noncytosol-targeted NUOT genes could elevate the average d*N* such that they appeared similar to that of the other gene groups in Geraniaceae. Furthermore, the *Silene* values represent rates for ribosomal proteins only whereas our study included sequences encoding diverse functions. That the d*N*/d*S* values for ribosomal proteins were also elevated in the slower evolving pair of *Silene* species, whereas d*S* values did not differ significantly, suggests that these genes may be more labile than other functional groups that were included in our data set. Alternatively, assuming the ratio of d*N* of different gene groups in Geraniaceae is the ancestral state for both families, the ratio of d*N* of the three gene groups in *Silene* could be caused by lineage-specific

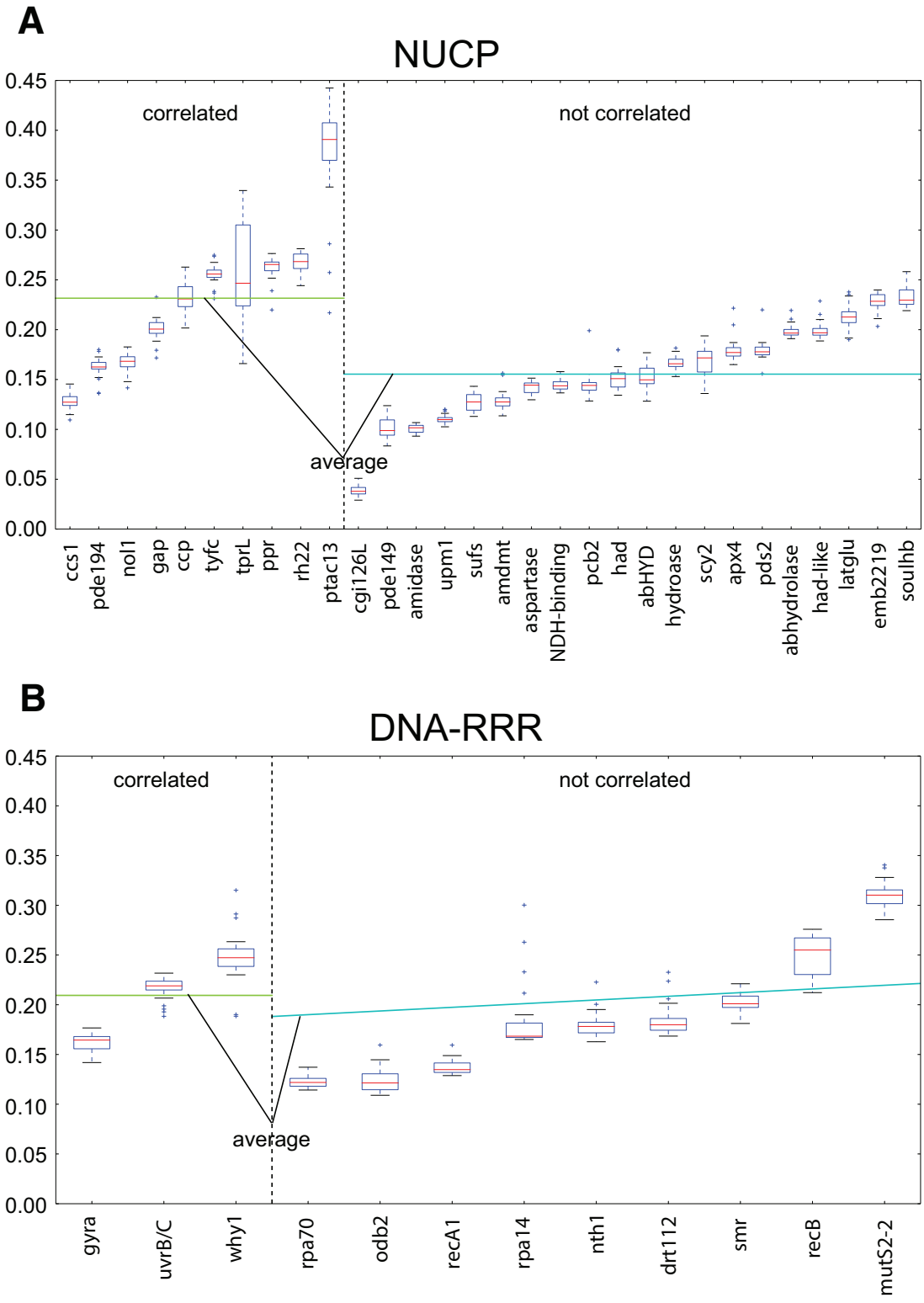

Fig. 4.—Comparison of d*N* of genes with and without significant correlation to genome complexity. d*N* of genes from (*A*) NUCP or (*B*) RRR gene groups were compared. NUCP, nuclear encoded plastid targeted control genes; RRR, plastid-targeted DNA replication, recombination and repair genes. Boxes represent sorted rates ranging from 25% to 75%, dashed lines extend to maximal and minimal rates, red lines represent the median rate for each gene group, and outliers are shown as "+." Horizontal lines represent average rate of gene groups with correlation (green) or without correlation (cyan).

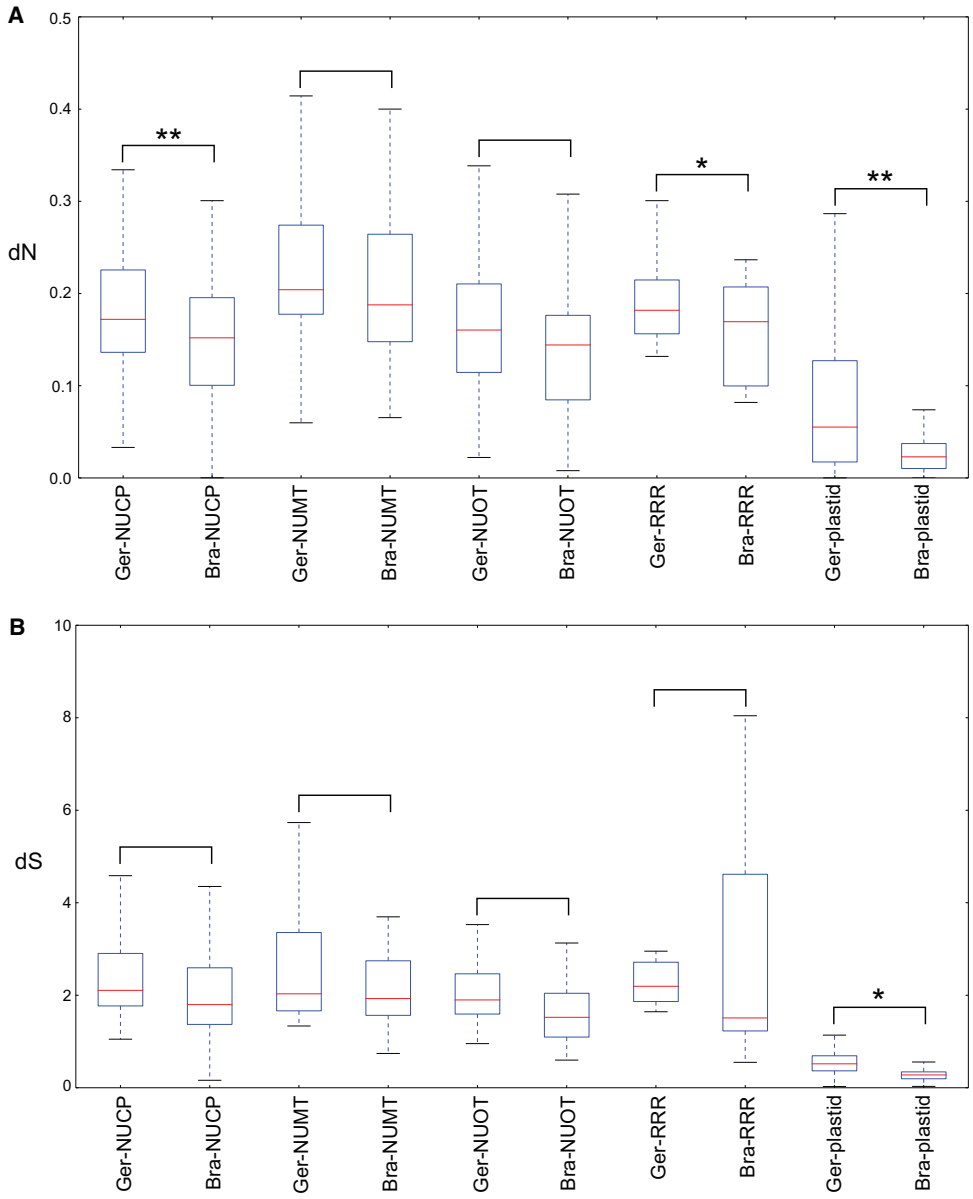

Fig. 5.—Comparison of evolutionary rates between gene groups in Geraniales and Brassicales. Ger, gene groups are from Geraniales; Bra, gene groups are from Brassicales; NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear encoded control genes; RRR, plastid-targeted DNA replication, recombination and repair genes; plastid, plastid-encoded genes. Boxes represent sorted rates ranging from 25% to 75%, dashed lines extend to maximal and minimal rates, red lines represent the median rate of each gene group. $*P < 0.05$ and $**P < 0.001$.

acceleration of d*N* in both plastid- and mitochondrial-targeted gene groups.

Correlation of Genome Complexity and Nucleotide Substitution Rates

Previous studies in Geraniaceae revealed highly elevated nucleotide substitution rates and extensive genome rearrangements (Chumley 2006; Blazier et al. 2011; Guisinger et al. 2011; Weng et al. 2014). One possible explanation for the high level of genome complexity is dysfunction of the nuclear-encoded plastid-targeted DNA-RRR genes (Guisinger et al. 2008; Weng et al. 2014). If this were the case, correlated changes in substitution rates of DNA-RRR genes and genome complexity would be predicted. Among the 12 plastid-targeted DNA-RRR genes investigated, d*N* of three genes (*gyrA*, *uvrB/C*, and *why1*) showed a significant correlation with the number of indels in protein and rRNA coding genes (CDR) and introns, and d*N* of the plastid genome (d*N* CP) (fig. 2; supplementary table S5, Supplementary Material online).

The earliest investigations of the nucleotide excision repair (NER) pathway reported excision and repair of DNA adducts in *Escherichia coli* following ultraviolet radiation exposure (Boyce and Howard-Flanders 1964; Pettijohn and Hanawalt 1964; Setlow and Carrier 1964). Years of study has revealed the ubiquity of the NER system among prokaryotes and, although more complex, a similar pathway has been described in detail for repair of DNA lesions in the eukaryotic nucleus where polymorphisms in pathway genes can impact cancer prognoses (Spitz et al. 2001; Qiao et al. 2002). Among eukaryotic organelles, mitochondria and the plastids of photosynthetic organisms were thought not to contain NER activity (Chen et al. 1996; Hada et al. 1998; Boesch et al. 2011). More recent investigations have suggested the presence of NER activity in animal mitochondria (Pohjoismäki et al. 2012) and the plastids of *Clamydomonas reinhardtii* (Vlcek et al. 2008) and *Arabidopsis* (Hays 2002). A putative protein-coding gene in *Arabidopsis* predicted to comprise both the uvrB/uvrC (UVR) motif and the hemimethylated DNA binding domain of *E. coli* YccV (syn. HspQ; Nishimura et al. 2013) was used to identify the Geraniaceae UvrB/C homolog in this study. Interestingly, the *E. coli* YccV also has a role in DNA-RRR as it is involved in the regulation of replication initiation protein DnaA expression (d'Alençon et al. 2003). The long history separating contemporary microorganisms from the endosymbiotic organelles of higher eukaryotes obscures evolutionary relationships that are sometimes detectable in gene sequences. However, although the UVR motif, involved in the binding of UvrC to the UvrB–DNA preincision complex (Moolenaar et al. 1995), is present in both *Es. coli* UvrB and UvrC, the hemimethylated DNA binding domain (yccV) is only present in UvrC suggesting a possible ancestry for the UvrB/C homolog identified among the Geraniaceae transcripts.

In addition to the correlation between d*N* of *uvrB/C* and plastome complexity, this study identified plastid-targeted homologs of genes encoding UvrD (DNA helicase II) in the nuclear transcriptomes of many species examined (supplementary data file S1, Supplementary Material online) supporting the notion that an NER-like system functions in Geraniaceae plastids.

The correlation between d*N* of *why1* and indels of CDR is supported by previous findings that *why1* knockouts accumulated plastid DNA duplications and deletions in *Arabidopsis*. In Whirly mutants reannealing of the nascent strand via microhomologies proximal to a stalled replication fork is proposed to reinitiate replication with the outcome determined by position at which the annealing occurs, upstream or downstream of the fork (Maréchal et al. 2009). Two issues may underlie the lack of correlation between *why1* and indels in introns and IG regions. First, sample size is much smaller for the intronic and IG regions with only 9 introns and 21 IGs compared with 63 coding regions. Second, alignment of plastid intron and IG regions is more problematic than coding regions (Graham et al. 2000; Shaw et al. 2007), which can make accurate estimate of the number of indels much more challenging.

The *gyrA* gene encodes subunit A of Gyrase, which is involved in DNA replication processes, including supercoiling or relaxing DNA, and concatenation or deconcatenation of DNA rings (Wang 1996; Levine et al. 1998; Singh et al. 2004). Although the influence of Gyrase on plastome nucleotide substitution rates is not clear, substitution rates could be affected by the DNA replication process in various ways (Tamura 1992; Stamatoyannopoulos et al. 2009; Liu et al. 2013).

Although a common factor could have increased both rates of nucleotide substitutions in the DNA-RRR genes (*uvrB/C*, *why1*, and *gyrA*) and the level of plastid genome complexity, the underlying connection between DNA-RRR genes and plastome complexity provides a biological basis for the correlation suggesting a more persuasive explanation. The dysfunction of DNA-RRR genes (*uvrB/C*, *why1*, and *gyrA*) could have facilitated the increase in genomic rearrangements, indels, and d*N* CP. Furthermore the former scenario does not explain why a common factor would affect only certain measures of genome complexity, that is, d*N* CP but not d*S* CP (synonymous substitution rates of the plastid genome) (supplementary table S5, Supplementary Material online).

Estimates of correlation between d*N* of DNA-RRR genes and plastome complexity included 90 nuclear encoded genes with three different subcellular locations as negative controls. Unexpected significant correlations were detected between genome complexity and both NUCP and NUMT genes (fig. 3). Two of the three NUMT genes that showed a correlation are dually targeted to plastids, suggesting that this result may be caused by the plastid component for two of the genes. Significant correlations with NUCP sequences involved genes with a wide diversity of functions (i.e., RNA binding, cytochrome c biogenesis, and cell cycle control; supplementary data files S1 and S3, Supplementary Material online) argue against the possibility that shared functional constraint maintains the correlation for the control genes. None of the DNA-RRR or the nuclear control genes showed correlation of both d*N* and d*S* with genome complexity, indicating that the correlations are not due to background mutation rates (fig. 2*B*). General d*N* acceleration may have caused the correlation of d*N* of NUCP and DNA-RRR genes with plastome complexity. However significant d*N* acceleration relative to noncorrelated genes was seen only for NUCP, not DNA-RRR genes (fig. 4). It could be that the accelerated rate in some NUCP sequences spuriously permitted their correlation with plastome complexity measures.

The generally accelerated nucleotide substitution rates in Geraniaceae could nonetheless be masking differences between gene groups. Comparison of rates for the same gene sets (NUCP, NUMT, NUOT, RRR, and plastid) between Geraniales and another, more conservatively evolving group could reveal potential causes of the unexpected correlations. Brassicales have highly conserved plastid genomes with a

single gene loss documented for one family (Ruhlman and Jansen 2014). Significant acceleration was found for d*N* in NUCP, DNA-RRR, and plastid-encoded gene groups in Geraniales relative to Brassicales (fig. 5*A*). The acceleration of d*N* in NUCP and DNA-RRR genes provides an explanation for the correlation of the d*N* of these gene groups with genome complexity. The lack of significant acceleration of d*S* for NUCP and DNA-RRR genes (fig. 5*B*) demonstrates that the correlation of d*N* is not due to differences in background mutation rates between the orders.

Unusually high levels genome rearrangement and elevated rates of nucleotide substitution make Geraniaceae an attractive system to study nuclear–organelle genome coevolution. Previous studies hypothesized that these phenomena could result from alterations in DNA repair and recombination mechanisms (Guisinger et al. 2008, 2011; Weng et al. 2014). The identification of significant correlations between nonsynonymous substitution rates of DNA-RRR genes and some measures of genome complexity, and the connection between the functions of DNA-RRR proteins and specific metrics of genome complexity support this hypothesis.

Acknowledgments

This work was supported by the National Science Foundation (IOS-1027259 to R.K.J. and T.A.R.) and from Vice President for Educational Affairs Professor Dr Abdulrahman O. Alyoubi at King Abdulaziz University (KAU), Jeddah, Saudi Arabia to J.Z., R.K.J., and J.S. The authors thank Anna Yu for helpful discussions on measures of plastid genome complexity, Scott Hunicke-Smith and Heather Deiderick of UT GSAF for assistance with Illumina sequencing, Texas Advanced Computing Center for supercomputer access, and the Plant Resources Center at UT for housing voucher specimens.

Supplementary Material

Supplementary figures S1–S3, tables S1–S7, and data files S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Bernt M, et al. 2005. CREx: inferring genomic rearrangements based on common intervals. Bioinformatics 23:2957–2958.

Blazier JC, Guisinger MM, Jansen RK. 2011. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol Biol. 76:263–272.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. Cell and molecular biology of plastids. Springer, p. 29–63.

Boesch P, et al. 2011. DNA repair in organelles: pathways, organization, regulation, relevance in disease and aging. Biochim Biophys Acta. 1813:186–200.

Boyce RP, Howard-Flanders P. 1964. Release of ultraviolet light-induced thymine dimers from DNA in *E. coli* K-12. Proc Natl Acad Sci U S A. 51:293–300.

Cai Z, et al. 2008. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol. 67:696–704.

Cappadocia L, et al. 2010. Crystal structures of DNA-Whirly complexes and their role in *Arabidopsis* organelle genome repair. Plant Cell 22:1849–1867.

Cappadocia L, Parent J-S, Sygusch J, Brisson N. 2013. A family portrait: structural comparison of the Whirly proteins from *Arabidopsis thaliana* and *Solanum tuberosum*. Acta Cryst Sect F Struct Biol Cryst Commun. 69:1207–1211.

Chen JJ, Jiang CZ, Britt AB. 1996. Little or no repair of cyclobutyl pyrimidine dimers is observed in the organellar genomes of the young *Arabidopsis* seedling. Plant Physiol. 111:19–25.

Cho HS, et al. 2004. DNA gyrase is involved in chloroplast nucleoid partitioning. Plant Cell 16:2665–2682.

Chumley TW, et al. 2006. The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol. 23:2175–2190.

Cosner ME, Raubeson LA, Jansen RK. 2004. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol Biol. 4:27

D' Alençon E, et al. 2003. Isolation of a new hemimethylated DNA binding protein which regulates *dnaA* gene expression. J Bacteriol 185:2967–2971.

Darling AE, Mau B, Perna NT. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147

Day A, Madesis P. 2007. DNA replication, recombination, and repair in plastids. In: Bock R, editor. Cell and molecular biology of plastids Springer, p. 65–119.

Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol. 49:827–831.

Duarte JM, et al. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol. 10:61

Fajardo D, et al. 2012. Complete plastid genome sequence of *Vaccinium macrocarpon*: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genet Genomes 9:489–498.

Gaut BS. 1998. Molecular clocks and nucleotide substitution rates in higher plants. In: Hecht MK, Macintyre RJ. Clegg MT, editors Evolutionary biology, Springer, p. 93–120.

Graham SW, Reeves PA, Burns ACE, Olmstead RG. 2000. Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. Int J Plant Sci. 161(6 Suppl):S83–S96.

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

Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci U S A. 105:18424–18429.

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

Hada M, Hashimoto T, Nikaido O, Shin M. 1998. UVB-induced DNA damage and its photorepair in nuclei and chloroplasts of *Spinacia oleracea* L. Photochem Photobiol 68:319–322.

Hays JB. 2002. *Arabidopsis thaliana*, a versatile model system for study of eukaryotic genome-maintenance functions. DNA Repair 1:579–600.

Jansen RK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A. 104:19369–19374.

Jansen RK, Ruhlman TA. 2012. Plastid genomes of seed plants. In: Bock R, Knoop V, editors. Genomics of chloroplasts and mitochondria, advances in photosynthesis and respiration, Springer, p. 103–126.
Kagale S, et al. 2014. Polyploid evolution of the Brassicaceae during the Cenozoic era. Plant Cell 26:2777–2791.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Knox EB. 2014. The dynamic history of plastid genomes in the Campanulaceae *sensu lato* is unique among angiosperms. Proc Natl Acad Sci U S A. 111:11097–11102.
Levine C, Hiasa H, Marians KJ. 1998. DNA gyrase and topoisomerase IV: biochemical activities, physiological roles during chromosome replication, and drug sensitivities. Biochim Biophys Acta. 1400:29–43.
Liu L, De S, Michor F. 2013. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nat Commun. 4:1502
Maréchal A, Brisson N. 2010. Recombination and the maintenance of plant organelle genome stability. New Phytol 186:299–317.
Maréchal A, et al. 2009. Whirly proteins maintain plastid genome stability in *Arabidopsis*. Proc Natl Acad Sci U S A. 106:14693–14698.
Martínez-Alberola F, et al. 2013. Balanced gene losses, duplications and intensive rearrangements led to an unusual regularly sized genome in *Arbutus unedo* chloroplasts. PLoS One 8:e79685
Milligan BG, Hampton JN, Palmer JD. 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol Biol Evol. 6:355–368.
Moolenaar GF, et al. 1995. The C-terminal region of the UvrB protein of *Escherichia coli* contains an important determinant for UvrC binding to the preincision complex but not the catalytic site for 3′-incision. J Biol Chem. 270:30508–30515.
Nishimura K, et al. 2013. ClpS1 is a conserved substrate selector for the chloroplast Clp protease system in *Arabidopsis*. Plant Cell 25:2276–2301.
Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 271:511–523.
Pazos F, Ranea JAG, Juan D, Sternberg MJE. 2005. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol. 352:1002–1015.
Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14:609–614.
Perry AS, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH. 2002. Evolutionary re-organisation of a large operon in Adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res. 9:57–162.
Pettijohn D, Hanawalt P. 1964. Evidence for repair-replication of ultraviolet damaged DNA in bacteria. J Mol Biol. 9:395–410.
Pohjoismäki JLO, et al. 2012. Oxidative stress during mitochondrial biogenesis compromises mtDNA integrity in growing hearts and induces a global DNA repair response. Nucleic Acids Res. 40:6595–6607.
Qiao Y, et al. 2002. Rapid assessment of repair of ultraviolet DNA damage with a modified host-cell reactivation assay using a luciferase reporter gene and correlation with polymorphisms of DNA repair genes in normal human lymphocytes. Mutat Res. 509:165–174.
Ruhlman TA, et al. 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biol. 15:100
Ruhlman TA, Jansen RK. 2014. The plastid genomes of flowering plants. Methods Mol Biol. 1132:3–38.
Sabir J, et al. 2014. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnol J. 12:743–754.
Schwarz EN, et al. 2015. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps16* in papilionoids. J Syst Evol. 5:458–468.
Setlow RB, Carrier WL. 1964. The disappearance of thymine dimers from DNA: an error-correcting mechanism. Proc Natl Acad Sci U S A. 51:226–231.
Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am J Bot 94:275–288.
Singh BN, Sopory SK, Reddy MK. 2004. Plant DNA topoisomerases: structure, function, and cellular roles in plant development. Crit Rev Plant Sci. 23:251–269.
Sloan DB, Triant DA, Wu M, Taylor DR. 2014. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Mol Biol Evol. 31:673–682.
Spitz MR, et al. 2001. Modulation of nucleotide excision repair capacity by XPD polymorphisms in lung cancer patients. Cancer Res. 61:1354–1357.
Stamatoyannopoulos JA, et al. 2009. Human mutation rate associated with DNA replication timing. Nat Genet. 41:393–395.
Tamura K. 1992. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol Biol Evol. 9:814–825.
Tesler G. 2002a. Efficient algorithms for multichromosomal genome rearrangements. J Comp Syst Sci. 65:587–609.
Tesler G. 2002b. GRIMM: genome rearrangements web server. Bioinformatics 18:492–493.
Vlcek D, Sevcovicová A, Sviezená B, Gálová E, Miadoková E. 2008. *Chlamydomonas reinhardtii*: a convenient model system for the study of DNA repair in photoautotrophic eukaryotes. Curr Genet. 53:1–22.
Wall MK, Mitchenall LA, Maxwell A. 2004. *Arabidopsis thaliana* DNA gyrase is targeted to chloroplasts and mitochondria. Proc Natl Acad Sci U S A. 101:7821–7826.
Wang JC. 1996. DNA topoisomerases. Annu Rev Biochem. 65:635–692.
Weng M-L, Blazier JC, Govindu M, Jansen RK. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol. 31:645–659.
Wolf PG. 2012. Plastid genome diversity. In: Wendel JF, Greilhuber J, Dolezel J, Leitch IJ, editors. Plant genome diversity volume 1. Springer, p. 145–154.
Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 84:9054–9058.
Xu Y-Z, et al. 2011. MutS HOMOLOG1 is a nucleoid protein that alters mitochondrial and plastid properties and plant response to high light. Plant Cell 23:3428–3441.
Yang Z. 2007. PAML4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zampini É, Lepage É, Tremblay-Belzile S, Truche S, Brisson N. 2015. Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in *Arabidopsis* and humans. Genome Res. 25:645–654.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.
Zhang J, Ruhlman TA, Mower JP, Jansen RK. 2013. Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. BMC Plant Biol. 13:228
Zhang J, Ruhlman TA, Sabir J, Blazier JC, Jansen RK. 2015. Coordinated rates of evolution between interacting plastid and nuclear genes in Geraniaceae. Plant Cell 27:563–573.

Associate editor: Bill Martin