Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation between Genome Rearrangements, Repeats, and Nucleotide Substitution Rates

Mao-Lun Weng,*[1] John C. Blazier,[1] Madhumita Govindu,[1] and Robert K. Jansen[1,2]

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

[2]Genomics and Biotechnology Section, Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

***Corresponding author:** E-mail: maolun@utexas.edu.

Associate editor: Michael Purugganan

Abstract

Geraniaceae plastid genomes are highly rearranged, and each of the four genera already sequenced in the family has a distinct genome organization. This study reports plastid genome sequences of six additional species, *Francoa sonchifolia*, *Melianthus villosus*, and *Viviania marifolia* from Geraniales, and *Pelargonium alternans*, *California macrophylla*, and *Hypseocharis bilobata* from Geraniaceae. These genome sequences, combined with previously published species, provide sufficient taxon sampling to reconstruct the ancestral plastid genome organization of Geraniaceae and the rearrangements unique to each genus. The ancestral plastid genome of Geraniaceae has a 4 kb inversion and a reduced, *Pelargonium*-like small single copy region. Our ancestral genome reconstruction suggests that a few minor rearrangements occurred in the stem branch of Geraniaceae followed by independent rearrangements in each genus. The genomic comparison demonstrates that a series of inverted repeat boundary shifts and inversions played a major role in shaping genome organization in the family. The distribution of repeats is strongly associated with breakpoints in the rearranged genomes, and the proportion and the number of large repeats (>20 bp and >60 bp) are significantly correlated with the degree of genome rearrangements. Increases in the degree of plastid genome rearrangements are correlated with the acceleration in nonsynonymous substitution rates (d*N*) but not with synonymous substitution rates (d*S*). Possible mechanisms that might contribute to this correlation, including DNA repair system and selection, are discussed.

Key words: **Geraniaceae, plastid genome, genome rearrangement, repeats, nucleotide substitution rates.**

Introduction

Plastid genome organization is highly conserved throughout seed plants, but a few unrelated lineages show extremely high levels of genomic upheaval (Jansen and Ruhlman 2012). Most plastid genomes have a quadripartite structure with two copies of inverted repeats (IRs) separated by small single copy (SSC) and large single copy (LSC) regions. Seed plant plastid genomes usually contain 101–118 distinct genes, with majority of these genes (66–82) coding for proteins involved in photosynthesis and gene expression, and with the remainder encoding transfer RNA (29–36) and ribosomal RNA (4) genes (Bock 2007; Jansen and Ruhlman 2012). Despite the high level of conservation in gene order and content across seed plants, extensive plastid genome rearrangements have been found in conifers (Hirao et al. 2008; Wu et al. 2009), Campanulaceae (Cosner et al. 2004; Haberle et al. 2008), Fabaceae (Cai et al. 2008), Geraniaceae (Chumley et al. 2006; Blazier et al. 2011; Guisinger et al. 2011), and Oleaceae (Lee et al. 2007).

The scale of genomic rearrangements in Geraniaceae is unprecedented, as four of the six genera, *Erodium*, *Geranium*, *Monsonia*, and *Pelargonium*, have highly rearranged yet distinct plastid genomes (Chumley et al. 2006; Blazier et al. 2011; Guisinger et al. 2011). These genomes differ greatly in IR size, gene order, gene and intron content, and gene duplications. IR size ranges from absent in *Erodium* (Blazier et al. 2011; Guisinger et al. 2011) to 76 kb in *Pelargonium hortorum* (Chumley et al. 2006). Although based on limited taxon sampling, three gene and intron losses (*trnT-GGU* and the introns of *rps16* and *rpl16*) were found to be shared across the family (Guisinger et al. 2011). The plastid genome of Geraniaceae is extremely rearranged, and the reconstruction of their evolutionary model based on insufficient taxa sampling was not possible. Greater taxon sampling is necessary to reconstruct the ancestral plastid genome organization in Geraniaceae and to infer genome rearrangement events leading to each genus.

Prevalence of repetitive DNA has been observed in highly rearranged plastid genomes (Chumley et al. 2006; Lee et al. 2007; Cai et al. 2008; Haberle et al. 2008; Guisinger et al. 2011). Plastid transformation studies demonstrated that genome recombination via small (7–14 bp, Kawata et al. 1997) or large (148 and 232 bp, Rogalski et al. 2006) repeats can cause inversions (IVs). Although several studies have suggested a correlation between the number and size of repeats and plastid genome rearrangements (Cai et al. 2008;

Article

Haberle et al. 2008; Guisinger et al. 2011), a rigorous statistical test of this correlation has not been performed. Geraniaceae is an ideal family to test this correlation because their plastid genomes show varying degrees of rearrangements.

In angiosperms, the nucleotide substitution rate of the plastid genome is lower than that of the nuclear genome but higher than that of the mitochondrial genome (Wolfe et al. 1987; Drouin et al. 2008). In Geraniaceae, particularly in *Pelargonium*, accelerated nucleotide substitution rates have been detected in mitochondrial and plastid genomes (Parkinson et al. 2005; Guisinger et al. 2008; Weng et al. 2012). Although nucleotide substitution and genomic rearrangement appear to be separate mutational phenomena, previous studies have identified a positive correlation between rates of nucleotide substitution and genomic rearrangements in γ-proteobacterial (Belda et al. 2005), insect (Shao et al. 2003), and arthropod (Xu et al. 2006) mitochondrial genomes. The correlation might imply a cause-and-effect relationship between nucleotide substitutions and genomic rearrangements, or a correlation resulting from another mechanism that accelerates both processes, such as an alteration in the DNA replication system (Xu et al. 2006). However, in the case of plastid genomes, this correlation has not been thoroughly investigated. One comparative study across angiosperms identified a positive correlation between branch lengths and the number of genomic rearrangement events, i.e., gene and intron losses and number of IVs (Jansen et al. 2007), but no studies directly estimated the rate of genomic rearrangements in plastids. Geraniaceae is an ideal group to test this correlation because it exhibits considerable variation in genomic organization and rates of nucleotide substitution.

In this study, we report complete plastid genomes for six species, three from Geraniaceae and three from other Geraniale families. These genomes, combined with previous published genomes, provide sufficient taxon sampling to reconstruct the ancestral plastid genome of Geraniaceae and to address four questions. 1) Is the ancestral plastid genome of Geraniaceae rearranged? 2) What rearrangement events occurred on the branches leading to each genus? 3) Are small or large repeats associated with genome rearrangements in Geraniaceae? 4) Is there a correlation between nucleotide substitution rates and frequency of genomic rearrangements?

Results

Genome Assembly

Due to the differences in read length and total number of reads, average coverage was different for the two sequencing platforms. The average genome coverage ranged from 2,096× to 4,490× for Illumina and 48× for 454 (supplementary table S1, Supplementary Material online). Because the DNA sample with enriched plastid DNA was used for the 454 sequencing, over 92% of the 454 raw reads assembled to plastid genome, whereas only 3.31–8.95% of the Illumina raw reads assembled to plastid genome (supplementary table S1, Supplementary Material online).

Genome Organization

Plastid genome size showed considerable variation within the three Geraniaceae species (table 1, supplementary figs. S1 and S2, Supplementary Material online). Except for *California macrophylla*, the genome sizes of the other five species were larger than the median genome size for land plant plastid genomes (The median genome size of 246 land plant plastid genomes available from NCBI, accessed April 1, 2013, was 154 kb). Among the six species, *P. alternans* had the largest genome (173,454 bp), *Hypseocharis bilobata* had the largest LSC (100,009 bp), and *Viviania marifolia* had the smallest SSC (4,551 bp; table 1).

The plastid genome of *H. bilobata* was missing two genes, *accD* and *trnT-GGU*, and three introns, *rpl16*, *rps16*, and *rpoC1* introns (supplementary fig. S1, Supplementary Material online). The four RNA polymerase genes, *rpoA*, *rpoB*, *rpoC1*, and *rpoC2*, usually found in two separate

Table 1. Comparison of Geraniales Plastid Genomes Sequenced in This Study.

	Melianthus villosus	*Francoa sonchifolia*	*Viviania marifolia*	*Hypseocharis bilobata*	*California macrophylla*	*Pelargonium alternans*
GenBank accession number	KF017614	NC_021101	KF240615	KF240616	JQ031013	KF240617
Size (bp)	156,510	157,312	157,291	165,002	149,202	173,454
LSC length (bp)	85,639	85,979	83,138	100,009	88,738	90,431
SSC length (bp)	17,773	18,318	4,551	6,743	15,856	6,795
IR length (bp)	26,549	26,509	34,801	29,125	22,304	38,114
Number of different genes	108	112	108	110	108	109
Number of different protein-coding genes (duplicated in IR)	74 (6)	78 (6)	74 (12)	77 (7)	75 (5)	76 (20)
Number of different tRNA genes (duplicated in IR)	30 (7)	30 (7)	30 (7)	29 (6)	29 (7)	29 (5)
Number of different rRNA genes (duplicated in IR)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)
Number of different genes with introns (two introns)	18 (3)	18 (3)	16 (3)	15 (3)	15 (2)	16 (3)
Percent of genome coding for genes	61.29	60.68	60.27	68.36	58.83	61.86
Gene density – total number of genes/genome length including IR (genes/kb)	0.80	0.82	0.83	0.78	0.83	0.80
GC content (%)	37.3	37.6	37.7	38.9	38.7	38.9

transcriptional units, have relocated into a single cluster in the LSC (supplementary fig. S1, Supplementary Material online). One tRNA gene, *trnQ-UUG*, had five duplicated copies scattered throughout the genome.

The plastid genome of *P. alternans* had an expanded IR, with the IR/LSC and IR/SSC boundaries shifted into *petD* and *ndhA*, respectively (supplementary fig. S1, Supplementary Material online). A highly divergent *accD* gene (36% identical to the *accD* gene in *Eucalyptus*) was found between *rbcL* and *psaI*.

The plastid genome of *C. macrophylla* had a 25-kb IV spanning from *trnQ-UUG* to *trnE-UUC* (supplementary fig. S1, Supplementary Material online). Three introns were missing, including *rpl16*, *rps16*, and the first intron of *clpP*. The two large hypothetical chloroplast open reading frame (*ycf*) genes, *ycf1* and *ycf2*, were also missing from *C. macrophylla*.

The plastid genome of *V. marifolia* had two nested IVs; an 8-kb IV spanning from *rbcL* to *trnT-UGU* and a 1.3-kb IV including *trnT-UGU* and *trnL-UAA* (supplementary fig. S2, Supplementary Material online). One ribosomal protein gene, *rps16*, and the first intron of *clpP* gene were missing. *accD* gene was highly divergent in the genome (34.1% identical to the *accD* gene in *Eucalyptus*). The two large *ycf* genes, *ycf1* and *ycf2*, were inferred to be pseudogenes due to the presence of internal stop codons.

The plastid genome of *Melianthus villosus* and *Francoa sonchifolia* were slightly larger than median size of land plant plastid genome (154 kb; table 1, supplementary fig. S2, Supplementary Material online). Plastid genome organization and gene content of *M. villosus* and *F. sonchifolia* were identical to the plastid genome of outgroups except that four NDH genes, *ndhD*, *ndhF*, *ndhH*, and *ndhK*, were pseudogenes in *M. villosus* due to the presence of one or more internal stop codons.

Ancestral Plastid Genome of Geraniaceae

To reconstruct the ancestral plastid genome of Geraniaceae, shared IVs and ancestral IR/SSC and IR/LSC boundaries were identified. The Mauve alignment showed extensive rearrangements and identified 37 locally collinear blocks (LCBs) shared by Geraniaceae and outgroups (supplementary fig. S3 and table S2, Supplementary Material online). Each LCB for 12 genomes was numbered from 1 to 37 and assigned a ± based on strand orientation (supplementary table S3, Supplementary Material online). Among these six synapomorphic characters identified by the comparison of the adjacency of LCBs (supplementary table S4, Supplementary Material online), the presence of adjacency (−10)(11) and the disruption of adjacency (9)(10) and (10)(11) were due an IV of LCB (10) in Geraniaceae. This IV converted the (9)(10)(11) order into (9)(−10)(11) in the Geraniaceae ancestral genome. LCB (10) corresponded to the 4-kb region spanning from *trnG-GCC* to *psbD* (supplementary table S2, Supplementary Material online). The presence of adjacency (35)(−37) and the disruption of adjacencies (35)(36) and (36)(37) were due to the change of IR/SSC boundary that moved *trnN-GUU* next to *ndhF* in Geraniaceae. Based on the most parsimonious reconstruction, the reduced SSC region shared by *Hypseocharis*, *P. hortorum*, *P. alternans*, and *Monsonia* would be the ancestral SSC in Geraniaceae (fig. 1). Because each of the six Geraniaceae species had a unique IR/LSC boundary that differs from other rosid species, the ancestral IR/LSC boundary in Geraniaceae would be the same boundary shared by other rosids based on the most parsimonious reconstruction. Overall, the reconstructed ancestral plastid genome of Geraniaceae had an IV between *trnG-GCC* and *psbD*, a reduced, *Pelargonium*-like SSC region and the same IR/LSC boundary as other rosids.

Genome Rearrangement

Comparison of the adjacency of LCBs in each genome showed a mixture of shared (synapomorphic), unique (autapomorphic), and homoplasious characters in Geraniaceae and Geraniales plastid genomes (supplementary table S4, Supplementary Material online). There were six and four synapomorphic characters shared by Geraniaceae and *Pelargonium*, respectively. The number of autapomorphic characters ranged from two in *California* and *Erodium carvifolium* to 19 in *Geranium*. For species with rearranged plastid genomes, *P. alternans* was the only one that did not have any autapomorphic characters. Comparison of the adjacency of LCBs also showed 20 homoplasious characters throughout the phylogeny.

The plastid genomes of *Hypseocharis*, *P. hortorum*, *P. alternans*, *Monsonia*, *Geranium*, *California*, *E. carvifolium*, and *E. texanum* were compared with the Geraniaceae ancestral plastid genome to identify genome rearrangement events for each genus. Genome rearrangement models that account for IR boundary shifts and IVs were proposed for each genus in Geraniaceae (supplementary figs. S4–S8, Supplementary Material online), and rearrangement events were plotted on the phylogenetic tree (fig. 2). The models estimated 15 IR contractions, eight IR expansions, and 46 IVs in Geraniaceae plastid genomes. Among the 46 IVs, those between *ycf3~psbZ*, *psaI~rps18*, and *trnE-UUC~atpI* were homoplasious events (fig. 2).

Thirty-seven LCBs were identified in the Geraniaceae by Mauve whole plastid genome alignment (supplementary fig. S3 and table S2, Supplementary Material online). The order of the 37 LCBs in each genome was number coded (supplementary table S3, Supplementary Material online) for estimating breakpoint (BP) and IV distances between genomes. The third type of genome rearrangement distance estimated the number of IR boundary shifts and IVs between two genomes (IRIV distance, see Materials and Methods). Table 2 shows the pairwise comparison of the three types of genome rearrangement distances. The largest estimated BP distance was 25 between *Geranium* and *Hypseocharis*. The largest estimated IV distance was 19 between *P. hortorum* and *Hypseocharis*. The largest estimated IRIV distance was 28 between *Geranium* and *Hypseocharis* and between *Geranium*

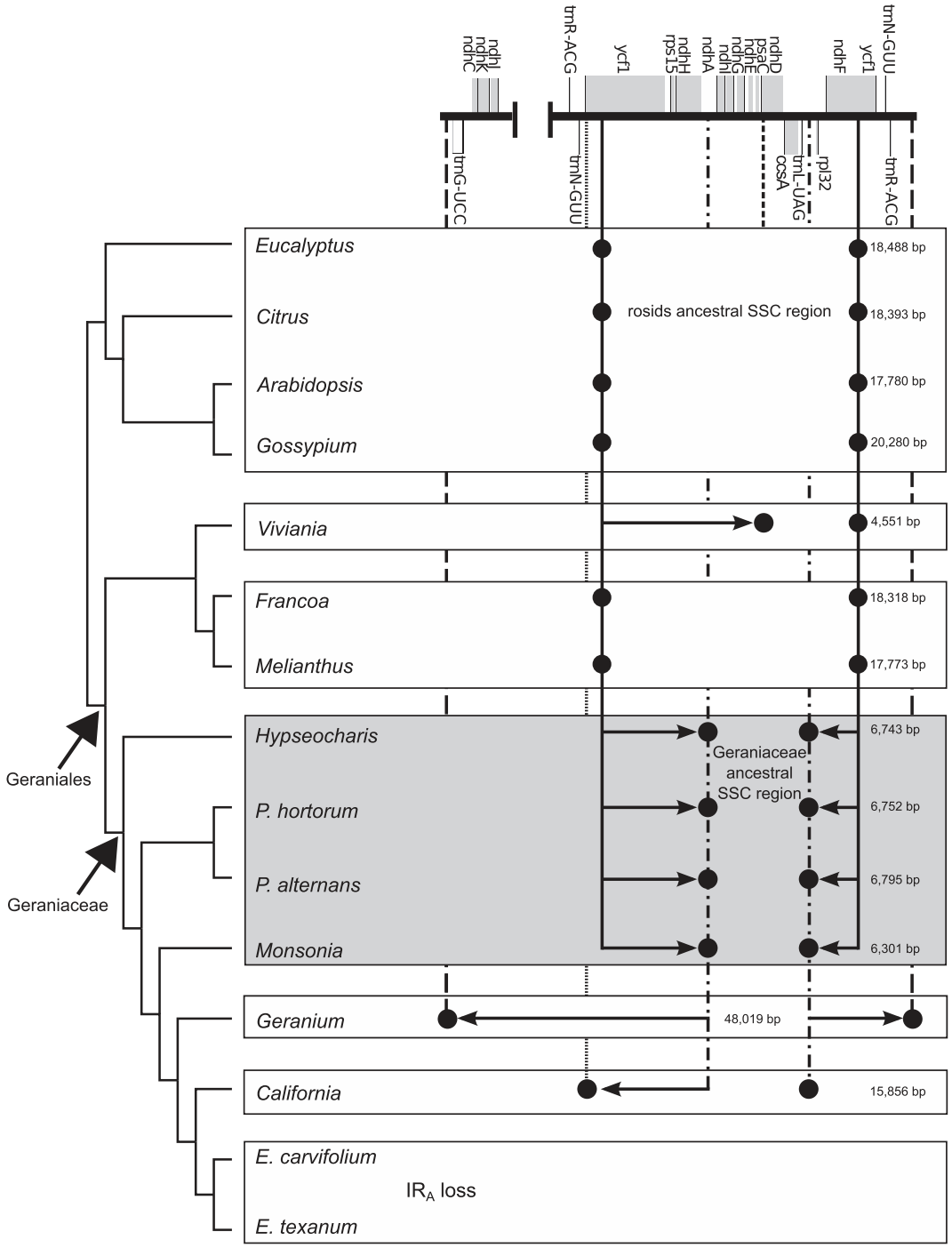


Fig. 1. Model of the rearrangements in the SSC region in Geraniaceae/Geraniales. Vertical lines denote the IR/SSC boundaries relative to the *Eucalyptus* genome shown on top. Dots on the solid vertical lines denote the ancestral IR/SSC boundary in rosids. Dots on the dashed lines denote the IR/SSC boundaries that deviate from the ancestral boundary in rosids. Arrows indicate the direction of IR/SSC boundary changes. *Hypseocharis*, *P. hortorum*, *P. alternans*, and *Monsonia* share the same SSC region highlighted in gray. This SSC region, including genes from *ndhA* to *trnL*, is the ancestral condition in Geraniaceae based on the most parsimonious reconstruction. This Geraniaceae ancestral SSC region was further expanded in *Geranium* and *California* independently. IRA was lost in *Erodium*.

and *E. texanum*. The three distances were highly correlated (BP~IV, $P < 0.001$, $r = 0.99$; BP~IRIV, $P < 0.001$, $r = 0.99$; IV~IRIV, $P < 0.001$, $r = 0.98$). The IRIV distance was used as the estimation of the degree of genome rearrangement in the later analyses because it did not exclude the IR region in the estimation.

Repeats in Geraniaceae and Geraniales Plastid Genomes

The number of repeats of different lengths was identified by BlastN searches of each plastid genome against itself (fig. 3A, supplementary table S5, Supplementary Material online).

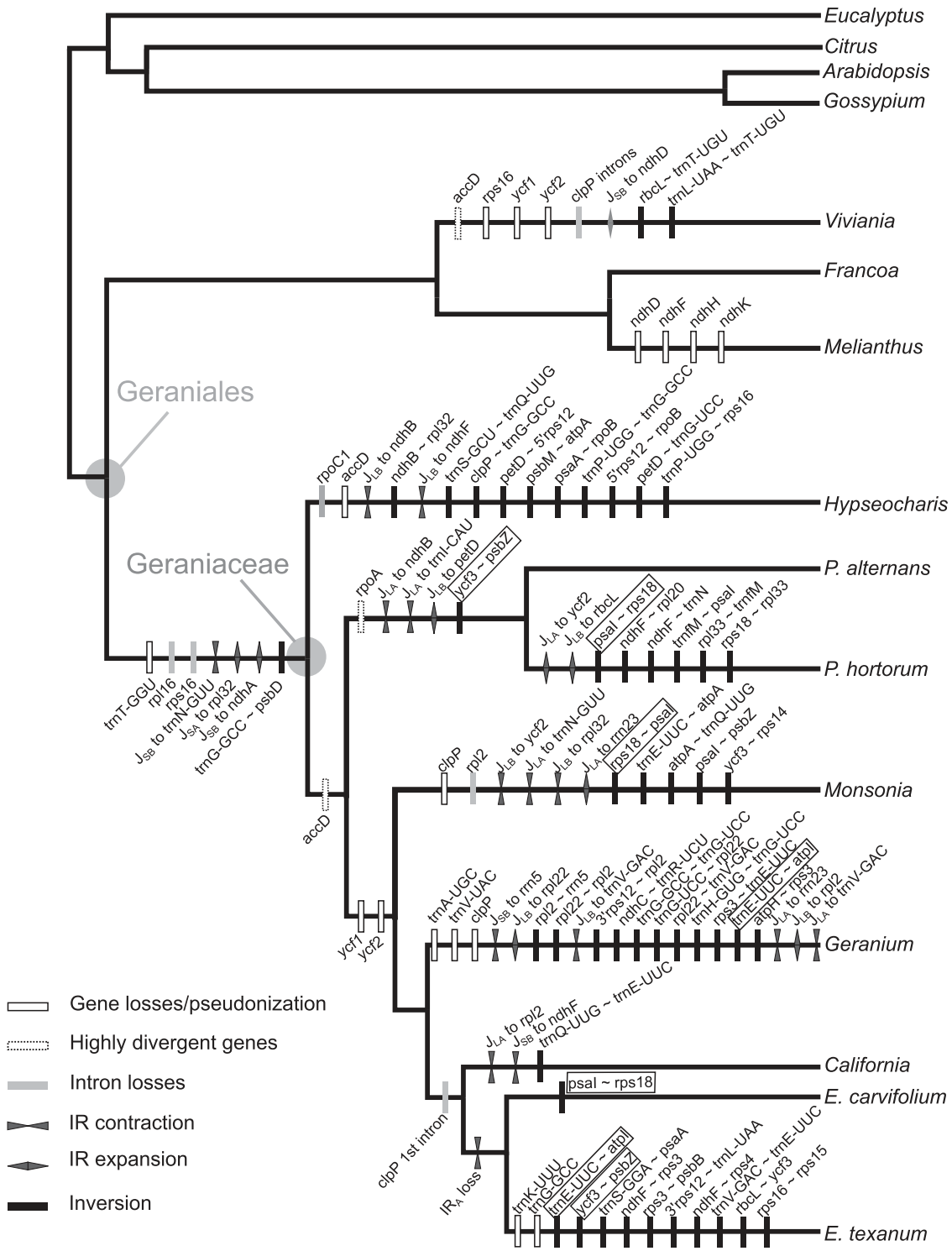


FIG. 2. Phylogeny of Geraniaceae and Geraniales based on 70 protein-coding genes with plastid genome rearrangement events mapped on the branches. Each node has 100% bootstrap support value.

Among the 15 plastid genomes compared, *Geranium palmatum* contained the largest number of repeats (505) while *F. sonchifolia* had the fewest (86). Repeats larger than 100 bp were absent from the plastid genomes of *Francoa* and *Melianthus*. Repeats larger than 300 bp were only present in Geraniaceae except *Hypseocharis*, and repeats larger than 1 kb, excluding the IR, were only present in three species, *P. hortorum*, *G. palmatum*, and *E. texanum*. The number of repeats varied among species from the same genus. The highly rearranged plastid genome of *E. texanum* contained 329 repeats but the relatively unrearranged *E. carvifolium* only had 136 repeats. The plastid genome of *P. hortorum* contained 293 repeats, whereas *P. alternans* had 216 repeats. The degree of genome rearrangement estimated by IRIV

Table 2. Pairwise Comparison of Genome Rearrangement Distances.

	Eucalyptus	*Viviania*	*Francoa*	*Melianthus*	*Hypseocharis*	*P. hortorum*	*P. alternans*	*Monsonia*	*Geranium*	*California*	*Erodium carvifolium*
Eucalyptus	—										
Viviania	3/2/3	—									
Francoa	0/0/0	3/2/3	—								
Melianthus	0/0/0	3/2/3	0/0/0	—							
Hypseocharis	16/13/16	19/15/19	16/13/16	16/13/16	—						
P. hortorum	15/12/15	17/14/18	15/12/15	15/12/15	21/19/23	—					
P. alternans	9/6/7	12/8/10	9/6/7	9/6/7	16/13/15	8/6/8	—				
Monsonia	14/11/13	16/13/16	14/11/13	14/11/13	20/16/21	15/12/20	13/11/12	—			
Geranium	17/12/20	20/14/23	17/12/20	17/12/20	25/17/28	24/18/27	17/12/19	23/17/25	—		
California	6/4/6	9/6/9	6/4/6	6/4/6	13/9/14	13/10/13	6/4/5	11/7/11	15/10/18	—	
E. carvifolium	7/4/6	9/6/9	7/4/6	7/4/6	15/11/14	12/8/13	7/4/5	10/7/11	16/10/18	4/2/4	—
E. texanum	15/13/16	16/15/19	15/13/16	15/13/16	20/16/24	19/15/23	14/11/15	17/14/21	22/17/28	12/11/14	13/11/10

NOTE.—The lower diagonal refers to BP/IV/IR shift and inversion (IRIV) distance.

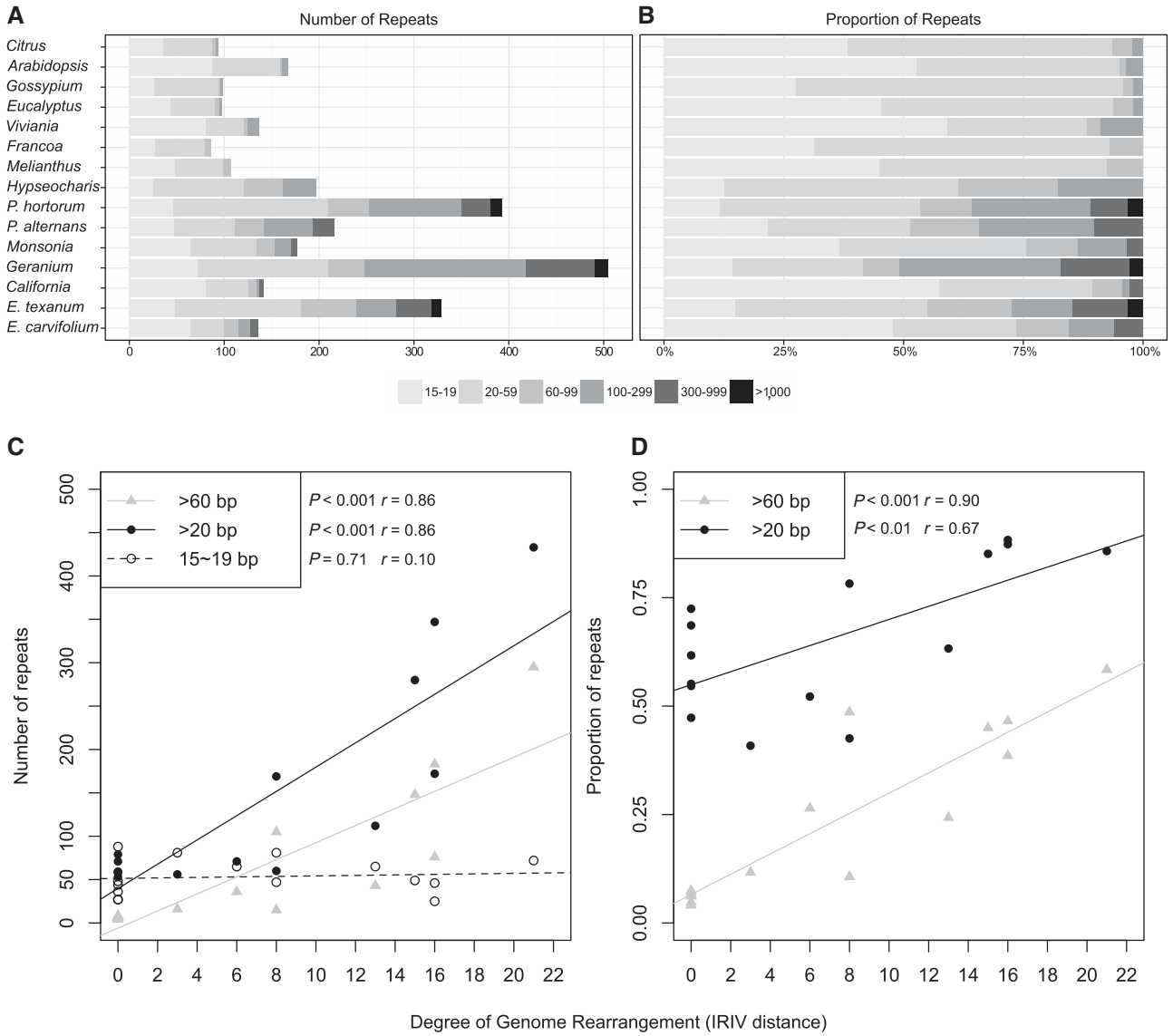


FIG. 3. Statistics of repeat analyses. (*A*) The number of different size repeats; (*B*) comparison of the proportion of different size repeats; (*C*) correlation of genome rearrangements with the number of repeats larger than 60 bp and 15–19 bp; and (*D*) correlation of genome rearrangements with the proportion of repeats.

distance was significantly correlated with the number of repeats larger than 60 bp ($P < 0.001$, $r = 0.82$) and 20 bp ($P < 0.01$, $r = 0.83$) but was not correlated with the number repeats smaller than 20 bp ($P = 0.83$; fig. 3*C*).

The proportion of repeats of different length among the total number of repeats identified was also computed (fig. 3*B*). In the unrearranged plastid genomes of rosids, *Melianthus*, and *Francoa*, over 90% of the repeats were shorter than 60 bp. In *Viviania* and *California*, over 80% of the repeats were shorter than 60 bp. The degree of genome rearrangement estimated by IRIV distance was correlated with the proportion of repeats larger than 20 bp ($P < 0.01$, $r = 0.69$) and 60 bp ($P < 0.001$, $r = 0.90$; fig. 3*D*).

The distribution of repeats was plotted on the plastid genomes (supplementary figs. S1 and S2, Supplementary Material online). The number of BPs in the genomes of *Viviania*, *Hypseocharis*, *P. alternans*, and *California* were identified in the Mauve alignment using the unrearrnaged *Melianthus* genome as a reference. The number of repeats located in the BP regions and non-BP regions of these four genomes were counted. The null hypothesis was that repeats not associated with BPs would be evenly distributed across the genome. The prediction would be that the number of repeats distributed in BP and non-BP regions is proportional to the size of these two regions. The alternative hypothesis was that repeats were aggregated in the BP regions. The χ^2 test showed that the number of repeats found in BP regions was significantly higher than expected by the null hypothesis (P value ranged from 8.9×10^{-4} in *California* to 2.2×10^{-16} in *Hypseocharis*; supplementary table S6, Supplementary Material online).

Correlation between Plastid Genome Rearrangements and Nucleotide Substitutions

Pairwise comparisons of d*N* and d*S* among 15 plastid genomes were computed using the codeml program in PAML and summarized in supplementary table S7, Supplementary Material online. The Pearson test showed that the degree of genome rearrangements estimated by IRIV distance was significantly correlated with nonsynonymous substitution rates (d*N*) ($P < 0.01$, $r = 0.3818$), d*N*/d*S* ratio ($P < 0.01$, $r = 0.4049$), but not with synonymous rates (d*S*) ($P = 0.59$, $r = 0.0680$) (fig. 4).

To address whether the elevated d*N* was caused by RNA editing, putative RNA editing sites were identified (supplementary material S1, Supplementary Material online), and d*N* and d*S* were reanalyzed based on edited sequences. The *t*-test showed that the mean of d*N* between original and edited sequences were not significantly different ($P = 0.664$), and the same correlation pattern was found between genome rearrangements and edited sequences (d*N*~IRIV, $P < 0.01$, $r = 0.3911$; d*S*~IRIV, $P = 0.58$, $r = 0.0685$). Two lines of evidence indicated that the d*S* was not saturated in the data set. First, the measured d*S* was all less than 0.5 (supplementary table S7, Supplementary Material online). Second, the regression analyses for linear and quadratic models did not indicate

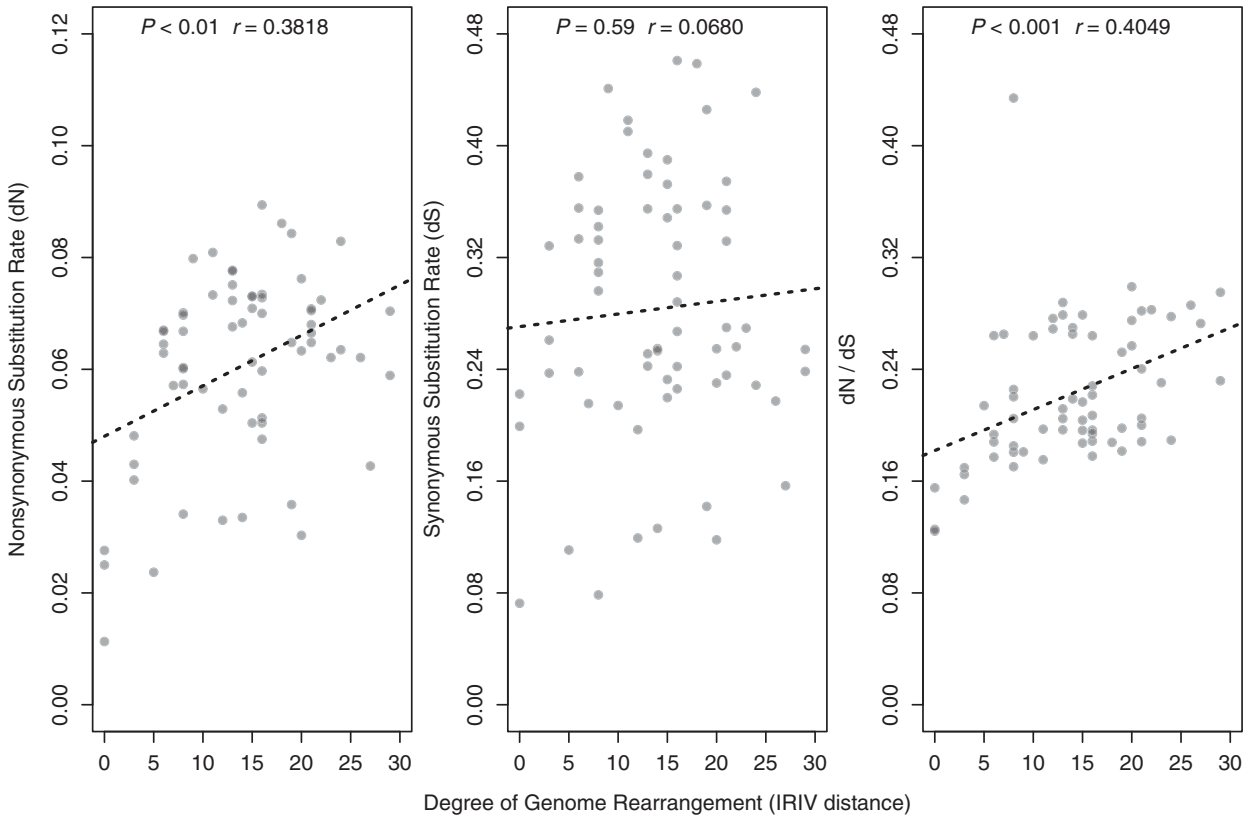


Fig. 4. Correlation of substitution rates and genomic rearrangements. Correlation of nonsynonymous (d*N*), synonymous (d*S*) substitution rates, and d*N*/d*S* ratio with the degree of genomic rearrangement (IRIV) distance.

saturation of dS ($R^2 = 0.636$ and = 0.655 for linear and quadratic model, respectively; supplementary fig. S9, Supplementary Material online). If dS was saturated in a d*N* versus dS scatter plot, a quadratic regression model with an increasing slope should fit the data significantly better than a linear model (Methods in Fares and Wolfe [2003]).

To assess whether the correlation between genome rearrangements and substitution rates was obscured by the shared phylogeny between pairwise comparisons, a Mantel test with 1,000 simulations and a phylogenetic independent comparison were performed. The Mantel test simulation involved randomizing rows and columns of the genome rearrangement matrix but holding the substitution rate matrix constant. Significance was assessed by comparing the observed correlation coefficient to a distribution of the correlation coefficients obtained from the simulations. The Mantel test confirmed that the degree of genome rearrangements was significantly correlated with the d*N* ($P < 0.01$) but not correlated with dS ($P = 0.28$). For the phylogenetic independent comparison, d*N*, dS, and the IRIV distance on corresponding branches in the phylogenetic tree were compared. Pearson correlation tests showed that IRIV distance was correlated with d*N* ($P < 0.01$, $r = 0.5551$) but not with dS ($P = 0.4723$, $r = 0.1416$) (fig. 5). The same comparison was performed with the control of time by estimating the absolute nonsynonymous (*rN*) and synonymous (*rS*) substitution rates on each branch in the phylogenetic tree. Pearson correlation tests showed that IRIV distance was correlated with *rN* ($P < 0.01$, $r = 0.5527$) but not with *rS* ($P = 0.1875$, $r = 0.2566$) (fig. 5).

Discussion

Plastid Genome Rearrangements in Geraniaceae Caused by IR Expansion/Contraction and IVs

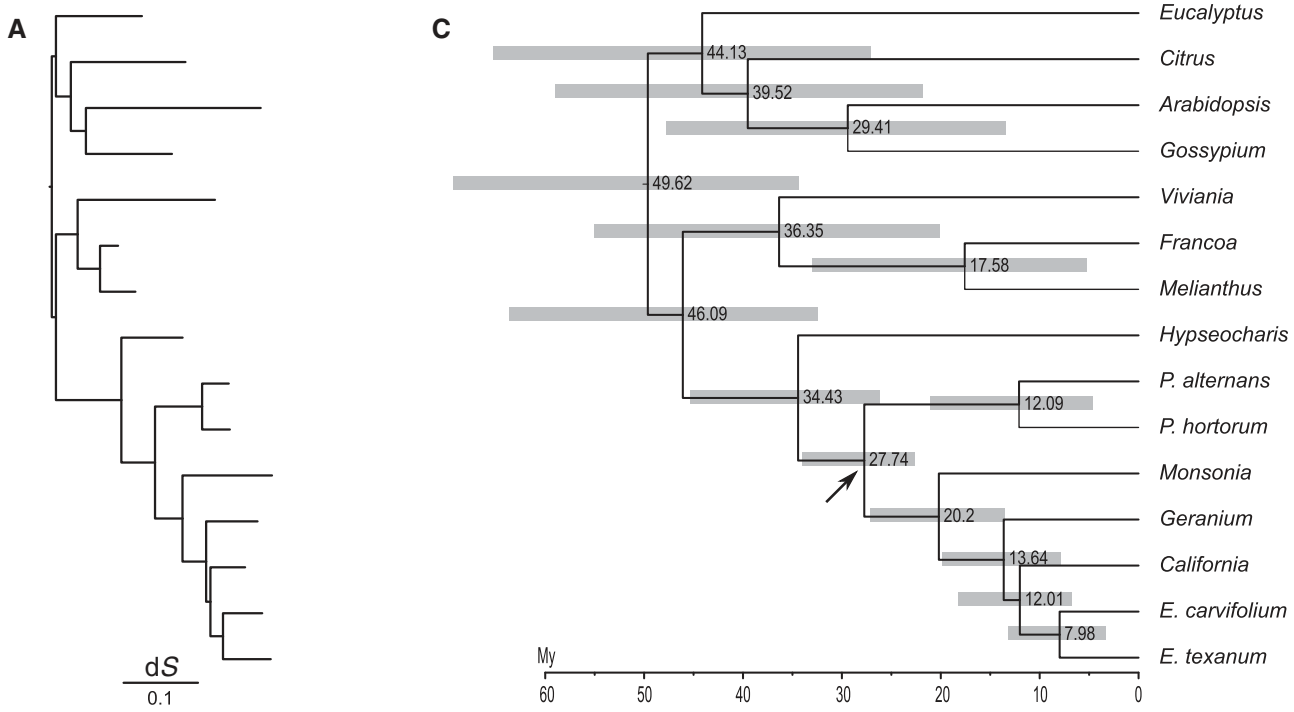
Although previous studies showed that the plastid genomes of four of the six genera in Geraniaceae, *Pelargonium*, *Geranium*, *Monsonia*, and *Erodium*, have experienced extensive genomic rearrangements (Chumley et al. 2006; Guisinger et al. 2011; Blazier et al. 2011), it was unclear whether these genomic changes were shared by all genera in Geraniaceae and Geraniales. Here we sequenced the plastid genomes of six additional species from the remaining two genera of Geraniaceae and three other Geraniales families to reconstruct the ancestral plastid genome in Geraniaceae and genome rearrangement events in each genus.

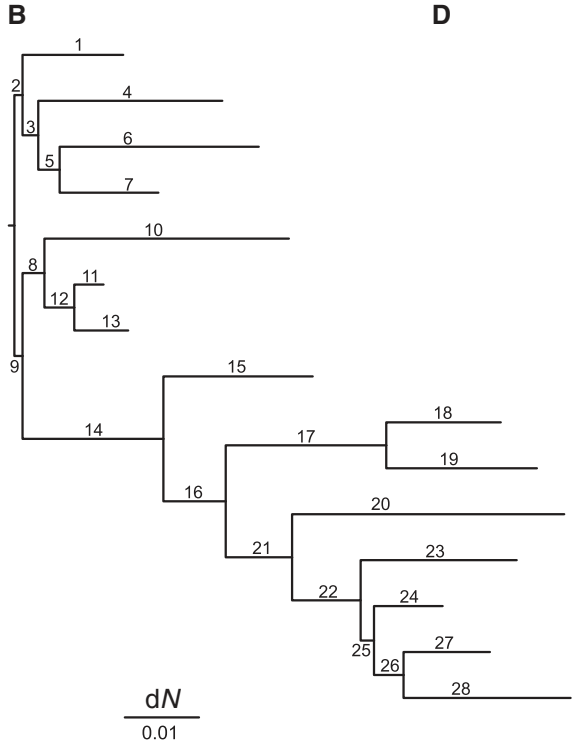
Our reconstruction models show few rearrangements from other rosids to the ancestral Geraniaceae plastid genome followed by a large number of independent rearrangements in each genus (fig. 2). Compared with the rosids, the ancestral Geraniaceae plastid genome has an IV between *trnG-GCC* to *psbD*, a *Pelargonium*-like SSC region and the same IR/LSC boundary as other rosids. A large number of autapomorphic characters and the absence of synapomorphic characters in the comparison of the adjacency of LCBs in each genome also indicate that independent rearrangements occurred in each genus (supplementary table S4, Supplementary Material online). In *Pelargonium*, the lack of autapomorphic characters in the LCB adjacency comparison (supplementary table S4, Supplementary Material online) indicates that the plastid genome in *P. alternans* likely represents the ancestral organization in the genus.

The reconstruction of the ancestral Geraniaceae plastid genome shows that IR expansion/contraction occurred repeatedly in the family (fig. 2). This resulted in dramatic IR size changes, ranging from absent in *Erodium* (Guisinger et al. 2011) to 75 kb in *P. hortorum* (Chumley et al. 2006). The size of the IR has also changed considerably within *Pelargonium*, where the IR ranges from 38 kb in *P. alternans* to 75 kb in *P. hortotum*. Although the degree of IR expansion and contraction in Geraniaceae is comparable to the situation in the green algae phylum Chlorophyta (Pombert et al. 2005, 2006) and gymnosperms (Wu et al. 2007; Lin et al. 2010, 2012), such dramatic IR boundary shifts within a single family and genus were unprecedented. Double-strand breaks in one of the two IRs followed by DNA repair using the other IR as template could be responsible for IR boundary shifts (Goulding et al. 1996).

IR expansion and contraction could cause gene duplications. Two cases of gene duplication, *petD* and *rpoA*, in *Pelargonium* are likely due to this phenomenon. In *P. alternans* the IR/LSC boundary expanded to *petD* producing a duplication of second exon of *petD*, which is a pseudogene (supplementary fig. S1, Supplementary Material online). This *petD* pseudogene is next to *trnL-CAA* in *P. hortorum* (supplementary fig. S5, Supplementary Material online). *rpoA* is highly divergent in *Pelargonium* compared with other angiosperms (Chumley et al. 2006; Guisinger et al. 2008). Three copies of *rpoA*-like ORFs are present in the *P. hortorum* plastid genome (Chumley et al. 2006), while only one copy of *rpoA* is in *P. alternans*. The position of *rpoA* in *P. alternans* next to the IR boundary suggests that duplications of *rpoA* in *P. hortorum* may be the result of a series of IR contractions/expansions.

IVs caused by recombination between repeated sequences are considered the main mechanism for changes in gene order in plastid genomes (reviewed in Jansen and Ruhlman [2012]). The number of reconstructed IVs in each Geraniaceae genus, except the IV between *trnG-GCC* to *psbD* that is shared by the family, ranges from 1 in *E. carvifolium* to 11 in *Geranium* (black bars in fig. 2). Although most of the IVs occur independently in each genus, three IVs are homoplasious events, *ycf3~psbZ*, *psaI~rps18*, *trnE-UUC~atpI* (fig. 2), suggesting there are some rearrangement hotspots in the genomes as observed in Campanulaceae (Cosner et al. 1997). Although most of the IVs in plastid genomes are in LSC region (Kim et al. 2005; Timme et al. 2007), the only IV that involves both LSC and IR regions is in ferns (Wolf et al. 2011). The plastid genome of *G. palmatum* has an extremely large SSC region, which includes several genes that are normally located in the LSC region (Guisinger et al. 2011). Two possible mechanisms could cause the expansion of SSC in *Geranium*. The first involves a loss of the IR followed by a replacement of IR in the middle of LSC region, resulting in a segment of the LSC moving to the SSC. The second involves an IV between the





D

Branch	IRIV	d*N*	d*S*	*rN*	*rS*
1	0	0.014	0.116	0.308 [0.208, 0.501]	2.620 [1.770, 4.261]
2	0	0.001	0.006	0.182 [0.138, 0.247]	1.020 [0.772, 1.383]
3	0	0.002	0.020	0.456 [0.334, 0.399]	4.338 [3.185, 3.802]
4	0	0.025	0.154	0.628 [0.420, 1.134]	3.899 [2.611, 7.046]
5	0	0.003	0.020	0.287 [0.256, 0.343]	2.008 [1.795, 2.400]
6	0	0.027	0.235	0.911 [0.562, 1.999]	7.987 [4.924,17.517]
7	0	0.013	0.116	0.452 [0.279, 0.992]	3.931 [2.423, 8.620]
8	0	0.003	0.029	0.308 [0.242, 0.348]	2.988 [2.347, 3.376]
9	0	0.001	0.006	0.312 [0.193, 0.576]	1.586 [0.984, 2.932]
10	3	0.033	0.185	0.905 [0.598, 1.639]	5.078 [3.354, 9.198]
11	0	0.004	0.024	0.222 [0.118, 0.743]	1.382 [0.735, 4.629]
12	0	0.004	0.030	0.213 [0.182, 0.270]	1.604 [1.369, 2.031]
13	0	0.007	0.048	0.415 [0.221, 1.391]	2.702 [1.437, 9.048]
14	4	0.019	0.088	1.630 [1.037, 3.006]	7.539 [4.798,13.908]
15	12	0.020	0.082	0.584 [0.443, 0.769]	2.393 [1.817, 3.151]
16	0	0.008	0.045	1.256 [0.738, 2.360]	6.771 [3.977,12.725]
17	4	0.022	0.063	1.380 [1.204, 1.678]	4.045 [3.528, 4.918]
18	8	0.020	0.037	1.679 [0.963, 4.366]	3.085 [1.769, 8.022]
19	0	0.015	0.036	1.274 [0.731, 3.312]	2.961 [1.698, 7.699]
20	9	0.037	0.120	1.817 [1.353, 2.715]	5.955 [4.434, 8.898]
21	0	0.009	0.037	1.180 [0.981, 1.305]	4.894 [4.068, 5.411]
22	0	0.009	0.032	1.402 [1.274, 1.631]	4.848 [4.404, 5.638]
23	17	0.021	0.069	1.540 [1.055, 2.665]	5.066 [3.471, 8.769]
24	3	0.009	0.047	0.774 [0.511, 1.388]	3.880 [2.560, 6.955]
25	0	0.002	0.006	1.104 [1.053, 1.525]	3.497 [3.333, 4.831]
26	1	0.004	0.017	0.993 [0.791, 1.177]	4.194 [3.340, 4.971]
27	1	0.012	0.053	1.454 [0.883, 3.515]	6.604 [4.011,15.970]
28	10	0.023	0.064	2.820 [1.712, 6.818]	8.045 [4.886,19.455]
P-value		**	0.472	** [**, *]	0.188 [0.216, 0.344]
r		0.555	0.142	0.553 [0.496, 0.456]	0.257 [0.242, 0.186]

Fig. 5. Branch-specific comparison between substitution rates and genomic rearrangements. (*A*, *B*) The maximum likelihood tree based on synonymous (dS) and nonsynonymous (d*N*) substitution rates for 70 protein-coding genes. Corresponding branches for correlation tests are labeled in (*B*). (*C*) The dated cladogram with median of node age estimates labeled. Arrow indicates the node with fossil age constraint using lognormal prior distribution. The gray bars indicate the 95% interval of the highest posterior density region (HPD) in age estimates. (*D*) The embedded table shows the correlation between d*N*, d*S*, *rN*, *rS*, and IRIV distances. *rN* and *rS* refers to the absolute nonsynonymous and synonymous substitution rates in subst/site/byr. Values in square brackets are estimates based on the 95% HPD intervals (see Materials and Methods). *P* values and correlation coefficient (*r*) of the Pearson correlation test between IRIV distance and substitution rates are shown in the bottom. Significance was marked with one ($P < 0.05$) or two ($P < 0.01$) asterisks.

IR and LSC regions that moves a segment of the LSC to the SSC. There is currently no evidence to support or eliminate either mechanism. A model for genome rearrangement in *Geranium* that involves an IV between the IR and LSC (supplementary fig. S7, Supplementary Material online) is more favorable because an IR loss and replacement has not been documented in any land plant plastid genome.

Gene order changes, caused by IV or IR boundary shifts, could bring genes that are normally separated in the plastid genomes next to each other. The clustering of four RNA polymerase genes in the *Hypseocharis* plastid genome (supplementary fig. S1, Supplementary Material online) is the only known case of these genes occurring together in published land plant plastid genomes. *rpoB*, *rpoC1*, and *rpoC2* are normally transcribed together in *rpoB* operon, whereas the *rpoA* gene is transcribed with other ten ribosomal protein genes in the *rpl23* operon. Although a simulation study suggested positive selection for clustering of functionally related genes on the plastid genome (Cui et al. 2006), we have no evidence that the four *rpo* genes in *Hypseocharis* are transcribed together. The fact that the *rpoB* and *rpl23* operons remain intact suggests that the clustering of four *rpo* genes may simply be a coincidence.

Gene and Intron Losses

Several gene and intron losses occur in Geraniales (fig. 2). The addition of plastid genome sequences for six species changes the findings of gene and intron losses reported by Guisinger et al. (2011). The duplication of *trnfM-CAU* shared by Geraniaceae is not as reported by Guisinger et al. (2011) because *P. alternans* only has one copy of *trnfM-CAU* (supplementary fig. S1, Supplementary Material online). It raises the possibility of independent duplications *of trnfM-CAU* in the family. The loss of *trnG-GCC* and the *trnG-UCC* intron reported by Guisinger et al. (2011) is due to the misidentification by DOGMA (Wyman et al. 2004). Except *E. texanum*, both *trnG-GCC* and *trnG-UCC* are present in all Geraniaceae included in this study based on two tRNA finding programs, tRNAscan-SE and ARAGORN. Gene losses or pseudogenizations in Geraniales include *rps16*, *ycf1*, *ycf2*, *clpP*, *ndh* genes, and four tRNAs (open bars in fig. 2). The intron losses throughout Geraniales include the genes *rpl16*, *rps16*, *clpP*, *rpoC1*, and *rpl2* (solid gray bars in fig. 2). These intron losses are not unique to Geraniales. The *clpP* first intron has been lost in the IR-lacking clade (IRLC), and in legumes (Jansen et al. 2008) and *clpP*, both introns have been lost in certain lineages in *Lychnis*, *Oenothera*, and *Silene* (Erixon and Oxelman 2008). The *rpoC1* intron is absent in some lineages in Poaceae (Katayama and Ogihara 1993), Passifloraceae, Fabaceae, Goodeniaceae (Downie et al. 1996), Cactaceae (Wallace and Cota 1996), and Aizoaceae (Thiede et al. 2007). The pseudogenization of four *ndh* genes in *Melianthus* is noteworthy because it is the only case in which all 11 *ndh* genes were not lost, suggesting a very recent loss; all 11 *ndh* genes are lost in Orchidaceae (Chang et al. 2006), *Welwitschia* (McCoy et al. 2008), a clade within *Erodium* (Blazier et al. 2011), and non-photosynthetic plants (Martin and Sabater 2010). The *accD* gene is so divergent in Geraniaceae and *Viviania* (open bars in fig. 2) that only the 3′ end is alignable with other rosids. A similar situation was observed in Fabaceae (Magee et al. 2010) and Poales (Harris et al. 2013). Recently, a functional transfer of truncated *accD* to the nucleus was found in Campanulaceae, suggesting that the 5′ end of the gene may not be important but only the catalytic domain at the 3′ end is critical (Rousseau-Gueutin et al. 2013).

Repeats and Genome Rearrangements

The number of repeats is correlated with the degree of genome rearrangement in green algal plastid genomes (Pombert et al. 2005, 2006). Previous studies have shown an association between repeats and genome rearrangement BPs (Lee et al. 2007; Guisinger et al. 2011), and plastid transformation studies demonstrated that IVs could be mediated by small (7–14 bp, Kawata et al. 1997) or large (148 and 232 bp, Rogalski et al. 2006) repeats. Association between repeats and genome rearrangement BPs (supplementary table S6, Supplementary Material online) and significant correlations between genome rearrangements and the proportion and the number of large repeats (>20 bp and >60 bp) (fig. 3) suggest that repeats larger than 20 bp may have facilitated rearrangements in Geraniales, although we cannot rule out the possibility that repeats may be introduced by DNA strand repair following a rearrangement event.

Correlation between Substitution Rates and Genomic Rearrangements

The correlation of d*N* and genomic rearrangements in Geraniaceae shows similarities to the situation in another angiosperm clade, *Silene* (Caryophyllaceae, Sloan et al. 2012). Both groups exhibit accelerated d*N* for only a subset of genes encoded in their plastid genomes; photosynthetic genes do not have accelerated rates of nonsynonymous substitutions while ribosomal proteins are highly accelerated. Although some of the most divergent genes in *Silene* have been lost in different lineages of the Geraniaceae (e.g., *accD*, *clpP*, *ycf1*, and *ycf2*), the genes that exhibit accelerated d*N* differ in the two groups. For example, the most rapidly evolving genes in Geraniaceae are the four plastid encoded polymerase genes (*rpoA*, *rpoB*, *rpoC1*, and *rpoC2*; Guisinger et al. 2008; Weng et al. 2012), but these genes are not highly accelerated in *Silene* (Sloan et al. 2012). Furthermore, the plastid genomes of Geraniaceae are much more rearranged than *Silene*, where rapidly evolving species have at most four IVs and only minor differences in the extent of IR (Sloan et al. 2012). In contrast, several lineages of Geraniaceae have experienced numerous IVs (up to 11) and extensive contraction and expansion of IR. Sloan et al. (2012) suggested that increased rates of sequence and structural evolution of plastid genomes in *Silene* may be associated with the process of functional gene transfer to the nucleus and positive selection.

Two possible explanations may underlie the observed correlation between d*N* and the degree of genome rearrangements in Geraniaceae plastid genomes: 1) the correlation was

due to a cause-and-effect relationship between elevated d*N* and genome rearrangements or 2) a common factor caused elevated d*N* and genome rearrangements. Shao et al. (2003) found a correlation between d*N* and genome rearrangements in insect mitochondrial genomes and proposed that high rates of nucleotide substitution led to genome rearrangements by introducing mutations in the replication initiation and termination sites. With regard to a shared causal factor for the two phenomena, mismatch repair system disruption could provide a plausible hypothesis. Mismatch repair systems recognize mismatches during the DNA replication process and prevent recombination between nonidentical sequences (Harfe and Jinks-Robertson 2000). Genes encoding MutS homologs, such as MSH1, have been identified in *Arabidopsis* (Culligan and Hays 2000; Maréchal and Brisson 2010) and the plastid genomes of MutS homolog mutants show rearrangements (Xu et al. 2011). The correlation between d*N* and genome rearrangements in Geraniaceae supports the notion that improper DNA repair may be responsible for both accelerated substitution rates and rearranged plastid genomes (Guisinger et al. 2008, 2011). However, defects in the DNA repair system should affect both synonymous and nonsynoymous sites. Therefore, it is puzzling that no significant correlation was found between d*S* and genome rearrangements. Although saturation of d*S* could have obscured the correlation with genome rearrangements, regression analyses did not indicate that saturation of d*S* had occurred (supplementary fig. S9, Supplementary Material online).

One possible explanation for the situation in Geraniaceae is that there are different repair mechanisms acting on genomic rearrangements versus nucleotide substitutions; therefore, the correlation may be spurious. DNA repair mechanisms in plastids are not well understood (Day and Madesis 2007; Ruhlman and Jansen, 2013). Alternatively, an improper DNA repair system followed by selection could explain the observed correlation between d*N* and genome rearrangements. A significant correlation between d*N*/d*S* ratio and genome rearrangements (fig. 4) suggests selection pressures on genome rearrangements though d*N*/d*S* ratios are all less than one. The DNA repair system in the Geraniaceae plastid genome may have been inefficient during the early diversification of the family, causing a burst of acceleration in nucleotide substitutions and genome rearrangements. Substitutions and rearrangements that generated faulty genomes were selected against and eliminated from the population, whereas substitutions and rearrangements that conferred an advantage might have been selected for. Although synonymous substitutions are relatively neutral, selection should have favored nonsynonymous substitutions that were beneficial, resulting in a correlation between d*N* and genome rearrangements. One possible advantage could be the selection for plastid genomes that are compatible with the nucleus. Nuclear-plastid incompatibility could impose selective pressure for the evolution of biparental inheritance as a mechanism to overcome incompatibility (Zhang and Sodmergen 2010). This argument is consistent with the fact that Geraniaceae have both biparental inheritance of plastids (Baur 1909; Metzlaff et al. 1981) and a nuclear plastid incompatibility system (Dahlgren 1925; Metzlaff et al. 1982). The role of nuclear-encoded, plastid-targeted genes in the evolution of the plastid genome of Geraniaceae, especially those genes involved in DNA replication, recombination, and repair, needs to be explored to test hypotheses regarding the bizarre evolutionary history of the organelle genomes in this family.

Materials and Methods

Taxon Sampling

Based on previously published phylogenies of Geraniaceae and Geraniales (Fiz et al. 2009; Palazzesi et al. 2012; Weng et al. 2012), six taxa were chosen: *H. bilobata*, *P. alternans*, and *C. macrophylla* from Geraniaceae and *F. sonchifolia* (Francoaceae), *M. villosus* (Melianthaceae), and *V. marifolia* (Vivianiaceae) from Geraniales. Four representative taxa from other rosids (*Citrus* NC_008334, *Arabidopsis* NC_000932, *Gossypium* NC_007944, *Eucalyptus* NC_008115), and five previously published (Chumley et al. 2006; Blazier et al. 2011; Guisinger et al. 2011) plastid genomes of Geraniaceae (*P. hortorum* NC_008454, *Monsonia speciosa* NC_014582, *G. palmatum* NC_014573, *E. texanum* NC_014569, *E. carvifolium* NC_015083) were also included.

DNA Extraction and Sequencing

Total genomic DNA of *F. sonchifolia*, *V. marifolia*, *H. bilobata*, *C. macrophylla*, and *P. alternans* was isolated from fresh leaf tissues using a modified version of hexadecyltrimethylammonium bromide (CTAB) procedure by Doyle JJ and Doyle JL (1987). Ground leaves were placed in CTAB buffer with 2% polyvinylpyrrolidone at 65 °C for 40 min. The solution was centrifuged and the supernatant was mixed with an equal volume of 24:1 chloroform:isoamyl alcohol and centrifuged twice. DNA was precipitated with isopropanol at −20 °C for an hour, pelleted by centrifugation, washed twice with 70% ethanol, and aspirated and resuspended in double-distilled water. The solution was then treated with RNase (Fermentas) at 37 °C for 1.5 h and the DNA was extracted using the same procedure above. DNA quantity and RNA contamination were verified with ethidium bromide on 1% agarose.

Plastid isolations for *M. villosus* used the sucrose gradient protocol described in Jansen et al. (2005). Plastid DNA was amplified by rolling circle amplification (Qiagen GmbH, Hilden, Germany) using bacteriophage Phi29 polymerase and random hexamer primers (Dean et al. 2001). Purity and quantity of plastid DNAs were verified by restriction enzyme digests, gel electrophoresis, and visualization of DNA fragments with ethidium bromide on 1% agarose gel.

Total genomic DNA was sequenced using Illumina HiSeq 2000 at Beijing Genomics Institute Corporation. For each species, approximately 60 million 100 bp paired-end reads were generated from a sequencing library with ~750 bp inserts. The plastid-enriched DNA from *M. villosus* was sequenced using 454 Titanium pyrosequencing at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois.

Genome Assembly and Annotation

Illumina paired-end reads were assembled de novo with Velvet v. 1.2.07 (Zerbino and Birney 2008) using a range of kmer sizes from 69 to 99. Nuclear and mitochondrial contigs were excluded using 1,000× and 2,000× coverage cutoffs. The assembly was performed on the Lonestar Linux Cluster from the Texas Advanced Computing Center (TACC). The 454 reads were assembled de novo using Newbler with default settings and using Mimicking Intelligent Read Assembly (MIRA) (Chevreux et al. 1999) with the accurate setting. The putative plastid contigs were identified using the mapping algorithm and assembled in Geneious v. 6.0.5 (Biomatters Ltd.). Genomes were annotated using Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al. 2004). Two *trnG* genes, *trnG-GCC* and *trnG-UCC*, that were misidentified by DOGMA were located by two programs, tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/, last accessed December 26, 2013) and ARAGORN (http://mbio-serv2.mbioekol.lu.se/ARAGORN, last accessed December 26, 2013). Circular and linear genome maps were drawn with OGDraw (Lohse et al. 2007).

Genome Rearrangement Analyses

Whole genome alignment among the 15 species was performed using the ProgressiveMauve algorithm in Mauve v 2.3.1 (Darling et al. 2010). A copy of IR was removed from plastid genomes with two IRs. The LCBs identified by the Mauve alignment were numbered to identify synapomorphic genome rearrangements and to estimate genome rearrangement distances. The synapomorphic genome rearrangements for Geraniaceae are defined as two or more adjacent LCBs that are shared by Geraniaceae but not by outgroups or vice versa. Three types of genome rearrangement distances were estimated. BP and IV distances were calculated using the web server of the Common Interval Rearrangement Explorer (Bernt et al. 2007). The third distance, IRIV, estimated the number of IR boundary shifts and IVs. This distance first counted the number of IR boundary shifts that allow two genomes to have equal gene content among single copy and IR regions. It then estimated the number of IVs between the SSC, LSC, and IR regions of the two genomes separately and found the optimal scenario transferring one genome into the other using Genome Rearrangements In Man and Mouse (GRIMM) (Tesler 2002).

Repeats Analyses

Repeats were identified using the command line version BlastN v. 2.2.26+ (word size = 16) by performing blast searches of each genome against itself. One copy of the IR was removed from plastid genomes with two IRs. The correlation between the number of repeats and the degree of genome rearrangement was tested using Pearson test in R v2.15 (R Development Core Team 2012). The association between repeats and genome rearrangement breakpoints was tested. BPs in the plastid genomes of *Viviania*, *Hypseocharis*, *P. alternans*, and *California* were identified by the Mauve alignment using the unrearranged *Melianthus* genome as a reference. A 2 kb region, 1 kb upstream and downstream from the BP, was defined as the BP region. The number of repeats located in the BP regions and non-BP regions of these four genomes were counted. Whether the distribution of repeats is proportional to the size of the BP and non-BP regions was tested using chi-square in R v2.15 (R Development Core Team 2012).

Phylogenetic Analysis and Estimation of Nucleotide Substitution Rates

Seventy protein-coding genes (supplementary table S8, Supplementary Material online) shared by all 15 taxa were extracted and aligned using Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh et al. 2002). Phylogenetic analysis was performed using maximum likelihood methods on the RAxML web server (Stamatakis et al. 2008). The maximum likelihood tree was then used as a constraint tree for plotting genome rearrangement events and estimating substitution rates. Nonsynonymous (d*N*) and synonymous (d*S*) nucleotide substitution rates were estimated using the codeml function in PAML 4.5 with codon frequencies determined by the F3 ×4 model (Yang 2007). Linear and quadratic regression models were compared to detect saturated synonymous substitutions (Methods in Fares and Wolfe [2003]). The correlation between the rate of nucleotide substitution and the rate of genome rearrangement was tested using Pearson test. Significance was assessed by generating 1,000 simulations in Mantel test using the ade4 package v1.4-17 (Chessel et al. 2011) in R v2.15 (R Development Core Team 2012). d*N*, d*S*, and IRIV distance from corresponding branches in the phylogenetic tree were extracted for phylogenetic independent comparisons.

Prediction of RNA Editing Sites

The prediction of RNA editing sites was performed using Predictive RNA Editor for Plants (PREP) (Mower 2009).

Estimation of Divergence Time and Absolute Substitution Rates

The divergence time was estimated using BEAST version 1.7.5 (Drummond et al. 2012) with one fossil calibration on the ancestral node of Geraniaceae excluding *Hypseocharis* (28.4 ± 0.1 My, references in Palazzesi et al. 2012). Two markers, *rbcL* and *matK*, with no accelerated substitution rates in Geraniaceae (Guisinger et al. 2008) were chosen for dating the phylogeny. The GTR + I + Γ substitution model was elected based on Akaike information criterion in jModeltest 2.1.4 (Darriba et al. 2012). A relaxed clock with lognormal distribution of uncorrelated rate variation was specified. A random starting tree with a Yule speciation process and a lognormal prior (mean = 16, log standard deviation = 0.2, offset = 13) on the single calibration node was adopted. Two independent Markov chains of 50,000,000 generations, sampled every 10,000th iteration, were generated. The adequate effective sample size (larger than 200) and convergence of the Markov chain Monte Carlo chains were diagnosed in Tracer v1.5 with 10% burn-in.

Absolute synonymous substitution rate (*rS*) was calculated by dividing a branch length in the dS tree by the time elapsed for that branch. The elapsed time for a branch was calculated as the difference in the median age estimates of nodes on both ends of that branch. Absolute nonsynonymous substitution rate (*rN*) was calculated in the same way.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Support was provided by the National Science Foundation (IOS-1027259 to R.J.K.). The authors thank Jeffery Mower and Randy Linder for helpful discussion, Tracey Ruhlman for comments on plastid inheritance and assistance on DNA isolation and for comments on an earlier version of the manuscript, Shane Merrell for maintaining the Geraniaceae plants at the greenhouse, Ian Gillespie for providing seeds of *C. macrophylla*, the Beijing Genomics Institute for Illumina sequencing, the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois for 454 sequencing, and the TACC at the University of Texas for access to supercomputers. The authors also thank two anonymous reviewers for the insightful comments on the manuscript.

References

Baur E. 1909. Das Wesen und die Erblichkeitsverhältnisse der 'Varietates albomarginatae hort' von *Pelargonium zonale*. *Z Indukt Abstammungs-Vererbungsl.* 1:330–351.

Belda E, Moya A, Silva FJ. 2005. Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. *Mol Biol Evol.* 22:1456–1467.

Bernt M, Merkle D, Ramsch K, Fritzsch G, Perseke M, Bernhard D, Schlegel M, Stadler P, Middendorf M. 2007. 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. Berlin (Germany): Springer. p. 29–63.

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK. 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.

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, et al. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol Biol Evol.* 23:279–291.

Chessel D, Dufour AB, Dray S. 2011. ade4: analysis of ecological data: exploratory and Euclidean methods in environmental sciences, version 1.4-17. Available from: http://cran.r-project.org/web/packages/ade4/index.html.

Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, p. 45–56. Available from: http://sourceforge.net/apps/mediawiki/mira-assembler/.

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

Cosner ME, Jansen RK, Palmer JD, Downie SR. 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr Genet.* 31:419–429.

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

Cui L, Leebens-Mack J, Wang LS, Tang J, Rymarquis L, Stern D, dePamphilis C. 2006. Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach. *BMC Evol Biol.* 6:13.

Culligan KM, Hays JB. 2000. *Arabidopsis MutS* homologs—AtMSH2, AtMSH3, AtMSH6, and a novel AtMSH7—form three distinct protein heterodimers with different specificities for mismatched DNA. *Plant Cell* 12:991–1002.

Dahlgren KVO. 1925. Die reziproken Bastarde zwischen *Geranium bohemicum* L. und seiner Unterart *deprehensum* Erik Almq. *Hereditas* 6:237–256.

Darling AD, Mau B, Perna NP. 2010. progressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. *PLoS One* 5:e11147.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 9:772.

Day A, Madesis P. 2007. DNA replication, recombination, and repair in plastids. In: Bock R, editor. Cell and molecular biology of plastids, Vol. 19. Berlin (Germany): Springer. p. 65–119.

Dean FB, Nelson JR, Giesler TL, Lasken RS. 2001. Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Gen Res.* 11:1095–1099.

Downie SR, Llanas E, Katz-Downie DS. 1996. Multiple independent losses of the *rpoC1* intron in angiosperm chloroplast DNA's. *Syst Bot.* 21:135–151.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* 19:11–15.

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.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29:1969–1973.

Erixon P, Oxelman B. 2008. Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP* gene. *PLoS One* 3:e1386.

Fares MA, Wolfe KH. 2003. Positive selection and subfunctionalization of duplicated CCT chaperonin subunits. *Mol Biol Evol.* 20:1588–1597.

Fiz O, Vargas P, Alarcón M, Aedo C, Garcia J, Aldasoro JJ. 2009. Phylogeny and historical biogeography of Geraniaceae in relation to climate changes and pollination ecology. *Syst Bot.* 33:326–342.

Goulding SE, Wolfe KH, Olmstead RG, Morden CW. 1996. Ebb and flow of the chloroplast inverted repeat. *Mol Gen Genet.* 252:195–206.

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.

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.

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.

Harfe BD, Jinks-Robertson S. 2000. DNA mismatch repair and genetic instability. *Annu Rev Genet.* 34:359–399.

Harris M, Meyer G, Vandergon T, Vandergon V. 2013. Loss of the acetyl-coA carboxylase (*accD*) gene in Poales. *Plant Mol Biol Rep.* 31:21–31.

Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. 2008. Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biol.* 8:70.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, 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, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, et al. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.* 395:348–384.

Jansen RK, Ruhlman TA. 2012. Plastid genomes of seed plants. In: Bock R, Knoop V, editors. Genomics of chloroplasts and mitochondria. Dordrecht (The Netherlands): Springer. p. 103–126.

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. 2008. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). *Mol Phylogenet Evol.* 48:1204–1217.

Katayama H, Ogihara Y. 1993. Structural alterations of the chloroplast genome found in grasses are not common in monocots. *Curr Genet.* 23:160–165.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

Kawata M, Harada T, Shimamoto Y, Oono K, Takaiwa F. 1997. Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs). *Curr Genet.* 31:179–184.

Kim KJ, Choi KS, Jansen RK. 2005. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Mol Biol Evol.* 22:1783–1792.

Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol Biol Evol.* 24:1161–1180.

Lin CP, Huang JP, Wu CS, Hsu CY, Chaw SM. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. *Genome Biol Evol.* 2:504–517.

Lin CP, Wu CS, Huang YY, Chaw SM. 2012. The complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction. *Genome Biol Evol.* 4:374–381.

Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW)—a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet.* 52:267–274.

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20:1700–1710.

Maréchal A, Brisson N. 2010. Recombination and the maintenance of plant organelle genome stability. *New Phytol.* 186:299–317.

Martin M, Sabater B. 2010. Plastid *ndh* genes in plant evolution. *Plant Physiol Biochem.* 48:636–645.

McCoy SR, Kuehl JV, Boore JL, Raubeson LA. 2008. The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. *BMC Evol Biol.* 8:130.

Metzlaff M, Börner T, Hagemann R. 1981. Variations of chloroplast DNAs in the genus *Pelargonium* and their biparental inheritance. *Theor Appl Genet.* 59:37–41.

Metzlaff M, Pohlheim F, Börner T, Hagemann R. 1982. Hybrid variegation in the genus *Pelargonium*. *Curr Genet.* 5:245–249.

Mower JP. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 37:W253–W259.

Palazzesi L, Gottschling M, Barreda V, Weigend M. 2012. First Miocene fossils of Vivianiaceae shed new light on phylogeny, divergence times, and historical biogeography of Geraniales. *Biol J Linnean Soc.* 107:67–85.

Parkinson C, Mower J, Qiu YL, Shirk A, Song K, Young N, dePamphilis C, Palmer J. 2005. Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. *BMC Evol Biol.* 5:73.

Pombert JF, Lemieux C, Turmel M. 2006. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. *BMC Biol.* 4:3.

Pombert JF, Otis C, Lemieux C, Turmel M. 2005. Chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. *Mol Biol Evol.* 22:1903–1918.

R Development Core Team. 2012. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/.

Rogalski M, Ruf S, Bock R. 2006. Tobacco plastid ribosomal protein S18 is essential for cell survival. *Nucleic Acids Res.* 34:4537–4545.

Rousseau-Gueutin M, Huang X, Higginson E, Ayliffe MA, Day A, Timmis JN. 2013. Potential functional replacement of the plastidic *accD* gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiol.* 161:1918–1929.

Ruhlman TA, Jansen RK. 2014. The plastid genomes of flowering plants. In: Maliga P, editor. Chloroplast biotechnology: methods and protocols. New York: Springer.

Shao R, Dowton M, Murrell A, Barker SC. 2003. Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. *Mol Biol Evol.* 20:1612–1619.

Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR. 2012. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. *Genome Biol Evol.* 4:294–306.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web-servers. *Syst Biol.* 75:758–771.

Tesler G. 2002. GRIMM: genome rearrangements web server. *Bioinformatics* 18:492–493.

Thiede J, Schmidt SA, Rudolph B. 2007. Phylogenetic implication of the chloroplast *rpoC1* intron loss in the Aizoaceae (Caryophyllales). *Biochem Syst Ecol.* 35:372–380.

Timme RE, Kuehl JV, Boore JL, Jansen RK. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *Am J Bot.* 94:302–312.

Wallace RS, Cota JH. 1996. An intron loss in the chloroplast gene *rpoC1* supports a monophyletic origin for the subfamily Cactoideae of the Cactaceae. *Curr Genet.* 29:275–281.

Weng ML, Ruhlman TA, Gibby M, Jansen RK. 2012. Phylogeny, rate variation, and genome size evolution in *Pelargonium* (Geraniaceae). *Mol Phylogenet Evol.* 64:654–670.

Wolf PG, Der JP, Duffy AM, Davidson JB, Grusz AL, Pryer KM. 2011. The evolution of chloroplast genes and genomes in ferns. *Plant Mol Biol.* 76:251–261.

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.

Wu CS, Lai YT, Lin CP, Wang YN, Chaw SM. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. *Mol Phylogenet Evol.* 52:115–124.

Wu CS, Wang YN, Liu SM, Chaw SM. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-

coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. *Mol Biol Evol.* 24: 1366–1379.

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

Xu W, Jameson D, Tang B, Higgs PG. 2006. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. *J Mol Evol.* 63:375–392.

Xu YZ, Arrieta-Montiel MP, Virdi KS, de Paula WB, Widhalm JR, Basset GJ, Davila JI, Elthon TE, Elowsky CG, Sato SJ, 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. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18: 821–829.

Zhang Q, Sodmergen. 2010. Why does biparental plastid inheritance revive in angiosperms? *J Plant Res.* 123:201–206.