Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants

Jennifer S. Hawkinsa,1, Stephen R. Proulx, Ryan A. Rappa, and Jonathan F. Wendela

aDepartment of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011; and bDepartment of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA 93106

Edited by Jeffrey L. Bennetzen, University of Georgia, Athens, GA, and approved August 18, 2009 (received for review April 21, 2009)

Transposable elements, particularly LTR-retrotransposons, comprise the primary vehicle for genome size expansion in plants, while DNA removal through illegitimate recombination and intramolecular homologous recombination serve as the most important counteracting forces to plant genomic obesity. Despite extensive research, the relative impact of these opposing forces and hence the directionality of genome size change remains unknown. In Gossypium (cotton), the 3-fold genome size variation among diploids is due largely to copy number variation of the gypsy-like retrotransposon Gorge3. Here we combine comparative sequence analysis with a modeling approach to study the directionality of genome size change in Gossypium. We demonstrate that the rate of DNA removal in the smaller genomes is sufficient to reverse genome expansion through Gorge3 proliferation. These data indicate that rates of DNA loss can be highly variable even within a single plant genus, and that the known mechanisms of DNA loss can indeed reverse the march toward genomic obesity.

C-value | cotton | genome evolution | Gossypium | transposable elements

Plants genomes vary enormously in size, from approximately 63 Megabases (Mb) in Genlisea margaretae (1) to greater than 120 Gigabases (Gb) in members of the Liliaceae (2, 3). The genesis of this extraordinary variation has been of interest for over half a century (4, 5), and numerous studies have shown that most genome size variation in plants can be ascribed to differences in the repetitive fraction of the genome, particularly long terminal repeat (LTR) retrotransposons (6–15). Additionally, transposable element (TE) proliferation is a dynamic process occurring repeatedly over short evolutionary timescales. For example, studies in maize suggest a doubling of its genome over as little as 3 million years due to TE accumulation alone (10, 11). The same pattern has been shown in Oryza australiensis, where three types of LTR-retrotransposons have proliferated recently and rapidly in episodic bursts that have doubled the genome within the last 2 million years (16). Similarly, copia- and LINE-like elements in some diploid member of Gossypium have amplified via episodic bursts within the last 5 million years, although at different times in each species’ evolutionary history (17).

Several mechanisms of DNA loss have been shown to attenuate genome expansion through TE proliferation. One is intramolecular homologous recombination, thought to occur predominantly between the directly repeated LTRs of retrotransposons, typically evidenced by a remaining solo LTR (12, 18). A second mechanism is illegitimate recombination, which generally takes place via nonhomologous end-joining (NHEJ) or slip-strand mispairing, resulting in small deletions (19, 20). Comparisons of internally deleted LTR-retrotransposons from rice and Arabidopsis suggest that illegitimate recombination may be the driving force behind DNA removal in these taxa with smaller genomes (19, 20). In these studies, however, the rate of genome size expansion through TE proliferation is greater than that of DNA removal, leading ultimately to larger genomes.

Given the rapid and recent accumulation of TEs in many plant genomes, combined with a short half-life for LTR-retrotransposons (20), insights into deletion dynamics and their impact on the directionality of plant genome size change are likely to emerge from studies of relatively recently diverged taxa (21). The cotton genus, Gossypium, is an especially good model in this respect. Gossypium (Malvaceae) is a monophyletic genus comprising approximately 50 diploid species of small trees and shrubs that are distributed throughout the world (22–25). Diploid members of this genus contain 13 chromosomes and are divided into eight (A–G, K) genome groups based on chromosome pairing behavior and interspecific fertility in hybrids (26, 27). Haploid nuclear content ranges 3-fold, from an average 885 Mb in the New World, D-genome species, to 2572 Mb in the Australian, K-genome species (28). This wide range in genome sizes and a well-established phylogeny make Gossypium an excellent model for studying the impact and dynamics of DNA removal as an evolutionary determinant of genome size.

Here, we focus on the abundant gypsy-like LTR-retrotransposon, Gorge3 (6). Using degenerate primers for the reverse transcriptase (RT) region of Gorge3, we amplified and performed phylogenetic analysis of 724 sequences from three Gossypium species that range 3-fold in genome size and from a phylogenetic sister group (24) to Gossypium, i.e., Gossypioides kirkii. Consistent with expectations from other studies in angiosperms, we show that recent episodic bursts of transposition have, in fact, occurred in each lineage, and that the magnitude of each burst is in direct positive correlation with genome size. In addition, however, we use a modeling approach to show that species with small genomes have experienced a faster rate of Gorge3 sequence removal relative to the rate of accumulation, leading to an overall decrease in genome size. The implication is that DNA removal is a powerful determinant of genome size variation among plants and that it can be a sufficiently strong force to not only attenuate, but reverse genome expansion through transposon accumulation.

Results

Phylogenetic Analysis and Timing of Transposition Events. A total of 724 unique reverse transcriptase (RT) sequences from G. herbaceum (A), G. raimondii (D), G. exiguum (K), and Gossypioides kirkii (outgroup) were subjected to phylogenetic analysis using neighbor-joining (Fig. 1). The resulting phylogeny contained two Gossypium-specific clades consisting of sequences from all three Gossypium species. Lineage-specific sequences from the A- and K-genome species, which have the larger genomes, formed


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be sent: Department of Genetics, University of Georgia, Athens, GA 30602. E-mail: jhawkins@uga.edu.
distinct clusters with short to medium branch lengths, while sequences from the D-genome and *G. kirkii* appeared to have longer branches. However, recent amplification of *Gorge3* even in the two species with small genomes, *G. kirkii* and *G. raimondii* (D-genome), was evidenced by small clusters with very short branch lengths present at the tips of multiple longer branches. Few nonlineage specific sequences were recovered from the taxa with larger genomes (*G. herbaceum* and *G. exiguum*).

Evaluation of the lineage-specific transpositional nature and timing of *Gorge3* in each genome revealed episodic bursts of activity since divergence from a common ancestor in all species, at different points in their respective evolutionary histories (Fig. 2). All A-genome pairwise comparisons among lineage-specific clades cluster at 95% sequence identity, suggesting a sudden burst of transposition approximately 2–3 million years ago, preceded and followed by relative quiescence. Similarly, the K-genome appears to have experienced a burst of transposition around 6.5 million years ago, but also appear to have experienced a burst very recent transpositional activity, perhaps within the last 2 million years. Similarly, *G. herbaceum* sequences clustered at 99% sequence identity, but also appear to have experienced a burst of transposition between 7 and 12 million years ago.

### Lineage-Specific Rates of DNA Gain Versus DNA Loss

Due to the inherent bias in PCR amplification toward younger intact sequences, *Gorge3* sequences from previously constructed random genomic shotgun sequence (GSS) libraries were used to quantitatively estimate the amount of *Gorge3* accumulation and deletion in each lineage. A second round of neighbor joining analysis was performed on the 724 PCR amplified sequences plus 294 *Gorge3* GSS sequences. The GSS sequences were partitioned into the three time periods (lineage-specific, *Gossypium* specific, and pre-*Gossypium*; see Methods) and the copy number and total number of *Gorge3* Mb originating at each time point was estimated (Table 1).

Surprisingly, *Gorge3* copy numbers were more abundant in, pre-*Gossypium* than at any other time point, for all taxa. Copy number estimates from this oldest time point in *G. kirkii* (3,001 ± 2,445) and the D-genome (4,731 ± 2,725) were not significantly different from one another, but many retained ancient copies of *Gorge3* were identified in the A-genome (22,272 ± 6,331) and twice as many ancient copies were recovered from the K-genome (43,037 ± 9,063). Copy number estimates for *Gossypium*-specific and lineage-specific time points were so low in the *G. kirkii* and D-genomes that they cannot be accurately estimated at this level of sampling. However, a consistent decrease in the number of copies originating during these two time periods is observed in the A- and K-genomes. Approximately 16,360 ± 5,434 *Gossypium*-specific and only 6,818 ± 3,515 lineage-specific copies were recovered from the A-genome. Similarly, 27,563 ± 7,271 *Gossypium*-specific and 12,089 ± 4,827 lineage-specific copies were identified in the K-genome. These copy numbers were subsequently used to estimate the total number of Mb from each time point in each genome, assuming the average *Gorge3* is 9.7 kb in length. While approximately 830 total Mb of *Gorge3* resides in the 2,460-Mb genome of *G. exiguum*, only 111 Mb originated specifically within the lineage. The same trend is observed in all of the genomes, with approximately 225 Mb pre-*Gossypium* and 60 Mb lineage-specific in the A-genome, and 50 Mb pre-*Gossypium* and only a few Mb lineage-specific in the D-genome.

### Modeling the Directionality of Genome Size Change

A simple growth model was used to infer changes in the rate of gain or loss of *Gorge3* across the phylogeny (see Methods). Based on the estimated *Gorge3* copy number in extant lineages (see above), the lineage-specific rates of gain or loss (indicated by numbers greater than or less than one, respectively) of *Gorge3* DNA (95% confidence intervals in brackets) are: A-genome: 3.91 [2.02, 6.12], D-genome: −4.96 [−7.58, −2.50], K-genome: 11.12 [8.83, 13.89], *G. kirkii* genome: −3.99 [−6.23, −2.70] (Fig. 3). The common ancestor was estimated to contain 28,878 copies of *Gorge3* (with a 95% confidence interval of [25,404–32,174]). Both A- and K-genomes are inferred to have undergone rapid expansion of *Gorge3*, while both the D-genome and *G. kirkii* have
experienced loss of Gorge3 DNA (Fig. 4). Thus, it appears that genome contraction through deletion of Gorge3 elements has played a dominant role in shaping the G. raimondii and Gospyoides kirkii lineages, whereas genome expansion through Gorge3 proliferation is implicated in the other two lineages.

**Discussion**

Here, we investigated both the quantitative and temporal nature of gypsy-like Gorge3 evolution in Gossypium. Previous results indicate that copy number variation of this particular LTR-retrotransposon family is primarily responsible for the 3-fold variation in genome size observed among diploid members of the genus (6). Congruent with these findings, we show here that Gorge3 has amplified differentially and independently in each of the lineages studied, with the highest copy number of sequences in the largest (K) genome and the lowest in the smallest (D) genome. However, the transpositional history in each lineage is distinctive and different. While lineage-specific transposition is episodic in nature in all genomes investigated, transpositional events occurred at different times in the evolutionary history of each clade. Episodic bursts of transposition have also been demonstrated in Oryza australiensis (16), a relative of rice with a large genome, and some diploid members of Gossypium (17), suggesting that episodic, transpositional bursts may be a general phenomenon in angiosperm evolution. To the extent that this pattern holds, it raises intriguing questions about the mechanisms that govern relatively long periods of evolutionary stasis, as well as the nature of the “triggers” that release TEs from suppression. Stress and interspecific gene flow are known to disrupt epigenetic regulation, and hence these factors may well be involved; in this respect it is noteworthy that Gossypium contains many documented examples of interspecific hybridization (33).

**Evidence for Genome Downsizing in Gossypium.** Comparisons between orthologous BACs from the A- and D-genomes have provided insight into the mechanisms and rates of DNA loss in Gossypium (32, 34). In a gene-rich region surrounding the CesA gene, both the genic and intergenic regions were highly conserved, but in the AdhA region this was not the case. Specifically,
in this region the A-genome contained unique transposable element insertions and the D-genome exhibited a 2-fold higher rate of indels, most containing hallmarks of illegitimate recombination, suggesting a higher rate of deletion in the smaller genome. Solo LTRs, indicative of DNA loss through intrastrand homologous recombination, were also evident, suggesting that both mechanisms are operating to remove DNA in *Gossypium*. Evidence presented here supports the interpretation that genome downsizing occurred in the D-genome lineage since its origin and despite TE proliferation. Our combined empirical and modeling approaches suggest that there is enormous lineage-specific variation in gain/loss ratio of *Gorse3* retrotransposons. The sequence data highlight an ancient and massive retrotranspositional event in the common ancestor of all *Gossypium* species as well as in the outgroup, *Gossypioides* *kirkii*. It is apparent that the A and K lineages have been unable to purge this ancient *Gorse3* DNA and have concomitantly accumulated more lineage-specific *Gorse3* copies. In contrast, the D-genome not only has discarded much of its ancient *Gorse3* complement, but has also suppressed other rounds of massive TE proliferation. Our modeling results highlight the robust nature of this conclusion. Under most reasonable scenarios, the gain/loss ratios are significantly biased for loss in the taxa having smaller genomes.

One caveat of the current study is that of sequence identification. The GSS sequences used to estimate copy numbers from each genome were identified via similarity searches, and if in fact the D- and *G. kirkii* genomes posses a higher rate of small deletions, then the more degenerate *Gorse3* sequences will be difficult to identify. This would lead to an underestimation of the total number of ancient copies residing in the smaller genomes. Every effort was made to avoid this potential pitfall by performing iterative blast searches within each GSS library to identify degenerate sequences with low sequence identity to *Gorse3*.

One may question whether the observed evolutionary trends for *Gorse3* are representative of the entire genome or if *Gorse3* is subject to evolutionary pressures unique to its particular genomic milieu. For example, gypsy-like retrotransposons have been shown to preferentially insert into pericentromeric heterochromatin in *A. thaliana* (36). It is unknown whether *Gorse3* exhibits similar insertion preferences or other biases, but the possibility remains that the inferences drawn here for rates of DNA loss and gain are not reflective of the genome overall. The veracity of our conclusions, both with respect to *Gossypium* and other angiosperms (and perhaps other eukaryotes), will only
studies may help elucidate the particular species-specific at-
forces induce or prevent TE proliferation? Further comparative
Why are some TE families more successful than others, even
is genome expansion differentially tolerated among lineages?
wherefores of these variations remain mysterious, however; why
on the fast track to a more streamlined genome. The whys and
little TE proliferation and, like
indicate that it “feasts” without purging over long evolutionary
against previously generated Gossypium GSS libraries (6) and 294 Gorge3 GSS sequences were recovered. The combined data set of 1,018 PCR and GSS sequences was subjected to neighbor-joining analysis and the approximate timing of the origin of each clade was determined (as described in 17). The number of GSS sequences originating within the three separate periods (lineage-specific, Gossypium specific, pre-Gossypium) was determined, and these numbers were used to estimate the copy numbers of Gorge3 from each period (6). The total number of Gorge3 Mb from each time was estimated by multiplying the copy number by 9.7 kb, the average length of Gorge3

Estimated Amount of Postspeciation Gorge3 Accumulation and Deletion. While PCR amplification of TEs is biased toward sequences that proliferated more recently, genomic shotgun sequences (GSS) comprise a putatively unbiased random sample of a genome. To quantitatively estimate the proportion of extant Gorge3 sequences originating at various evolutionary times and to circumvent PCR amplification bias, Gorge3 PCR sequences were queried against previously generated Gossypium GSS libraries (6) and 294 Gorge3 GSS sequences were recovered. The combined data set of 1,018 PCR and GSS sequences was subjected to neighbor-joining analysis and the approximate timing of the origin of each clade was determined (as described in 17). The number of GSS sequences originating within the three separate periods (lineage-specific, Gossypium specific, pre-Gossypium) was determined, and these numbers were used to estimate the copy numbers of Gorge3 from each period (6). The total number of Gorge3 Mb from each time was estimated by multiplying the copy number by 9.7 kb, the average length of Gorge3

While PCR amplification of TEs is biased toward sequences that proliferated more recently, genomic shotgun sequences (GSS) comprise a putatively unbiased random sample of a genome. To quantitatively estimate the proportion of extant Gorge3 sequences originating at various evolutionary times and to circumvent PCR amplification bias, Gorge3 PCR sequences were queried against previously generated Gossypium GSS libraries (6) and 294 Gorge3 GSS sequences were recovered. The combined data set of 1,018 PCR and GSS sequences was subjected to neighbor-joining analysis and the approximate timing of the origin of each clade was determined (as described in 17). The number of GSS sequences originating within the three separate periods (lineage-specific, Gossypium specific, pre-Gossypium) was determined, and these numbers were used to estimate the copy numbers of Gorge3 from each period (6). The total number of Gorge3 Mb from each time was estimated by multiplying the copy number by 9.7 kb, the average length of Gorge3

We wished to ascertain whether specific portions of the lineage (i.e., internodes) experienced biased gain/loss ratios of Gorge3. While there are many quantitative models of transposable element proliferation, we were most interested in describing gross changes in copy number. To this end, we applied a simple model for the change in Gorge3 copy number across the Gossypium phylogeny; for each lineage, the number of Gorge3 elements was assumed to follow exponential growth or decay (because each copy has the same chance of being deleted or transposed) with a fixed lineage-specific growth parameter. For example, the number of Gorge3 elements in the D-genome species is given by

\[ N_D = N_{\text{ADK}} e^{\lambda_D t} \]

where \( N_0 \) is the extant copy number in the D-genome species, \( N_{\text{ADK}} \) is the copy number in the common ancestor of the A, D, and K genome species, \( \lambda_D \) is the rate at which Gorge3 copy number changes and \( t \) is the amount of evolutionary time between the common ancestor and the D-genome species. Since we possess estimates for the Gorge3 copy number in extant lineages, we were able to reconstruct the internodes using least-squared parsimony and minimizing the squared change in Gorge3 copy number along each internode and tree branch. This means that the rate of Gorge3 copy number change along an internode is the average of three rates (the ancestor of that node and the two descendant internodes). Thus, for our estimated number of extant Gorge3

become clear following comparable studies in other genera, using multiple sequence types and with experimental designs that include sampling a diversity of taxa whose phylogenetic relationships are reasonably well known.

Conclusion

The present study demonstrates that genome size is a dynamic feature of plant genomes, even among recently diverged taxa within a single genus. As expected, some genomes appear to have a “one-way ticket to genomic obesity” (37), such as \( G. \ exiguum \), where a low deletion rate and highly proliferative suite of gypsy elements indicate that it “feasts” without purging over long evolutionary timeframes. On the opposite extreme, \( G. \ raimondii \) tolerates little TE proliferation and, like \( Arabidopsis \) and rice, seems to be on the fast track to a more streamlined genome. The whys and wherefores of these variations remain mysterious, however; why is genome expansion differentially tolerated among lineages? Why are some TE families more successful than others, even among closely related taxa? What internal factors and external forces induce or prevent TE proliferation? Further comparative studies may help elucidate the particular species-specific attributes that allow or prevent surplus transposition leading to genome expansion.

Methods

Plant Material, PCR Amplification, and Phylogenetic Analysis. Total genomic DNA from the A-genome species \( G. \ herbaceum \) (UMS; 1C = 1,667 Mb), the D-genome species \( G. \ raimondii \) (JPW; 1C = 880 Mb), the K-genome species \( G. \ exiguum \) (Gos 5184; 1C = 2,460 Mb), and the phylogenetic outgroup \( Gossypio-

\[ \text{Probability distributions for estimated parameters} \]

\[ \text{Rate of change in Gorge3} \]

\[ \text{Ancestral estimate of Gorge3 copy number} \]

\[ \text{Fig. 4. Probability distributions for estimated parameters. We generated 1,000 replicate datasets for the current Gorge3 copy number by sampling uniformly between our 95% confidence intervals. (A) The frequency distribution of exponential rate parameters is shown for the four genomes. Orange: \( G. \ kirkii \), Light blue: D, Green: A, Blue: K. Rates above 0 indicate that the number of Gorge3 elements is increasing, while numbers below 0 indicate that they are decreasing. In no case does the distribution overlap a rate of 0. (B) The frequency distribution of estimated number of Gorge3 elements in the common ancestor of the four genomes.} \]
copies we have a system of linear equations that can be solved for the rate of Gorge3 change along each branch of the tree:

\[
0.416r_g + \log(N_{GADK}) = 8.613
\]

\[
0.0687r_g + 0.257r_d + 0.0229r_u + 0.0229r_k
\]

\[
+ \log(N_{GADK}) = 9.068
\]

\[
0.714r_g + 0.0714r_d + 0.207r_u + 0.0309r_k
\]

\[
+ \log(N_{GADK}) = 10.783
\]

\[
0.714r_g + 0.071r_d + 0.031r_u + 0.147r_k
\]

\[
+ \log(N_{GADK}) = 11.391
\]

where \( r \) is the rate of Gorge3 change along the terminal branch leading to genome i, and \( N_{GADK} \) is the Gorge3 copy number in the common ancestor of all four species. This system of four equations has five unknowns (the four rates and the ancestral Gorge3 copy number). We defined a score function as the total amount of change in Gorge3 copy number along the phylogeny. Ancestral copy number was estimated by minimization of the score function. All of our growth rates are scaled to the standardized branch length.

While there are errors associated with our estimates of extant Gorge3 copy number, we performed the reconstruction 1,000 times by sampling uniformly within our 95% confidence intervals. In each case, the sign of the estimated growth rate was the same as that based on the estimates themselves.

ACKNOWLEDGMENTS. We thank Jordan Swanson for computational help, Mark Ungerer for providing primer sequences, Ashley Davidson for generation of sequence data, and Corrine Grover for providing molecular clock data. This work was supported by the National Science Foundation Plant Genome Program.