

**Science**

 AAAS

**Selection on Major Components of Angiosperm Genomes**

Brandon S. Gaut, *et al.*  
*Science* **320**, 484 (2008);  
DOI: 10.1126/science.1153586

***The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of July 8, 2008):***

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/320/5875/484>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/320/5875/484#related-content>

This article **cites 31 articles**, 16 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/320/5875/484#otherarticles>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

## PERSPECTIVE

# Selection on Major Components of Angiosperm Genomes

Brandon S. Gaut\* and Jeffrey Ross-Ibarra

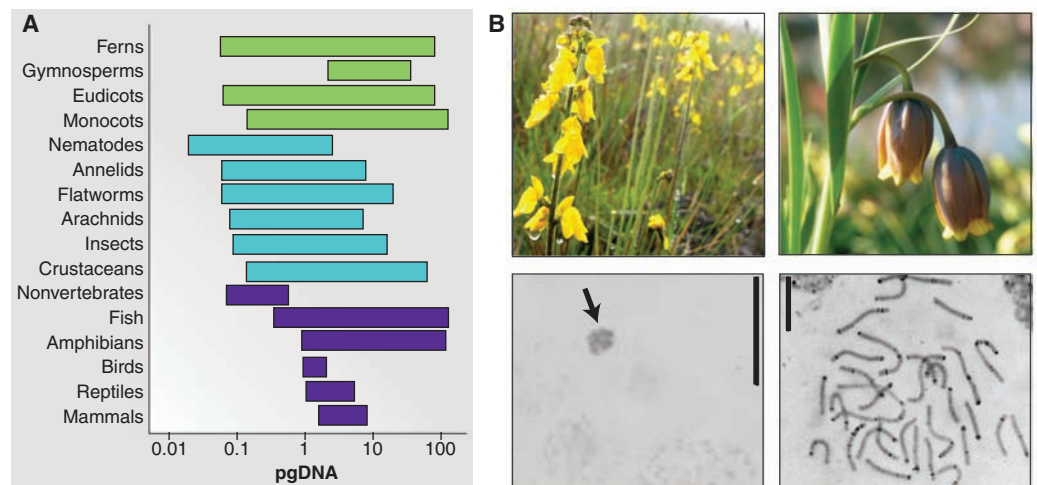
Angiosperms are a relatively recent evolutionary innovation, but their genome sizes have diversified remarkably since their origin, at a rate beyond that of most other taxa. Genome size is often correlated with plant growth and ecology, and extremely large genomes may be limited both ecologically and evolutionarily. Yet the relationship between genome size and natural selection remains poorly understood. The manifold cellular and physiological effects of large genomes may be a function of selection on the major components that contribute to genome size, such as transposable elements and gene duplication. To understand the nature of selection on these genomic components, both population-genetic and comparative approaches are needed.

Flowering plants are relative newcomers to the evolutionary stage, appearing for the first time 150 to 200 million years ago. Angiosperms have since radiated across the globe, quickly becoming a dominant life form on the planet. Mirroring their rapid diversification, the size of angiosperm genomes has changed rapidly as well: Higher plants vary ~2000-fold in genome size, from the 64-Mb genomes of *Genlisea* (corkscrew plants) (1) to the 124-Gb genomes of *Fritillaria* (the fritillary lilies) (Fig. 1) (2). Still, the nature of the relationship between genome size and natural selection is not well understood.

Genome size correlates with broad-scale patterns of plant biology. Plant species with large genomes tend to have large cells and large seeds, factors that are associated with a number of life-history traits. But plants with large genomes also have lower photosynthetic rates, grow more slowly, and are underrepresented in extreme environments (3). The ecological costs incurred by large genome size have a parallel evolutionary cost: Plant genera with the largest genomes tend to have the lowest species diversity, suggesting that genome size affects speciation rates (3).

Genome size also varies among individuals within a species, and such variation has been linked to selection. Individuals with the same chromosome number can vary as much as 40% in genome size (4). This intraspecific

diversity correlates with environmental clines and growth characteristics (5) and may also respond to indirect selection on other traits (6). However, the mechanisms that connect genome size to phenotype remain unclear. One possibility is that DNA content affects cell volume and



**Fig. 1. (A)** Despite being among the most recent of the groups depicted, monocots and eudicots encompass the widest range of genome sizes. Data are from (1, 2). pgDNA, picograms of DNA. **(B)** Photos depict the flower and metaphase squashes of *Genlisea* (1) on the left and a triploid *Fritillaria* (30) on the right. The arrow indicates the dividing *Genlisea* nucleus. Scale bars represent 10  $\mu$ m. [Photo credit: Fernando Rivadavia (*Genlisea*) and Christine Skelmersdale (*Fritillaria*)]

replication, leading to generally lower growth rates. The accrual of DNA may also have functional effects via gene regulation or copy-number variation (5). In any case, the manifold cellular and physiological effects of a larger genome may result in direct selection either on genome size itself or on the major components that contribute to genome size.

The largest contributor to genome size is repetitive DNA, particularly transposable elements (TEs). In fact, it is common for the majority of a plant's genome to consist of transposon-derived

DNA (Fig. 2). Much has been learned about TEs from genomic sequence data, including their distribution among species, their genomic locations of accumulation, the mechanisms by which they are purged from genomes, and their rates of proliferation. The last may be particularly impressive; the rice genome has increased >2% in size over the past few hundred thousand years because of TE activity alone (7). Individual animal genomes may also be element-rich (8), but plant genomes appear to vary more rapidly with respect to their transposon-derived component.

Given the apparently detrimental consequences of a large genome, it follows that the accumulation of transposons is probably deleterious to plant fitness. Many individual transposon insertions—such as those into coding regions—may be strongly deleterious, leading to their rapid loss from the gene pool (9). However, understanding the role of natural selection in shaping transposon diversity ultimately requires a population-genetic approach. In *Drosophila*, humans, *Arabidopsis*, and pufferfish, quantitative estimation of the strength of selection from population-genetic data suggests that TE insertions are on average slightly dele-

rious (10, 11) and thus expected to be purged from populations by natural selection. But the effectiveness of selection against TEs depends on the composite parameter  $N_e s$ , which includes not only the strength of selection  $s$  but also the effective population size  $N_e$ . Even if selection is relatively strong, species with low  $N_e$  may not be able to prevent transposons from accumulating within their genome. The proliferation of elements within plant genomes may thus reflect low  $N_e$  as much as low  $s$  (12). Though there have been surprisingly few studies of plant TE population genetics, this

Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California Irvine, Irvine, CA 92697-2525, USA.

\*To whom correspondence should be addressed. E-mail: bgaut@uci.edu

approach could go a long way toward illuminating the selective forces acting on plant genomes.

Population-genetic approaches have also been useful for identifying adaptive transposon insertions (13). This and other evidence suggests that TEs are not just deleterious but also contribute to genome function. In plants, transposons have been domesticated to become functional genes (14), have inserted complete exons into expressed genes (15), and have facilitated the formation of previously unrecognized genes via reverse transcriptase (16). Transposons are also potential sources of cis-regulatory elements and small RNAs. Moreover, transposable element activity can also accelerate the response to selection, presumably by producing genetic variation on which selection can act. An example is selection on bristle number in *Drosophila melanogaster*, where *p* element lines responded rapidly to selection but lines without active *p* elements did not (17). This study suggests that natural and artificial selection on a phenotypic trait could drive correlated increases in transposon activity in a manner antagonistic to selection against large genome size.

Gene duplication is another major contributor to plant genome size. The angiosperms sequenced to date contain more gene duplicates than animals (Fig. 2). Much of this duplication is due to polyploid events, which create complete genetic redundancy by copying every gene in the genome. Although many duplicated genes are lost as they accumulate mutations and deletions, this process is nonrandom (18). Genes related to transcription, signal transduction, and development are more likely to be retained as duplicates than other functional gene categories. This biased retention may result from variable sensitivity of genes to dosage effects, with selection acting to maintain proper stoichiometric ratios (18). Retention biases can also be taxon specific, perhaps explaining the high abundance of aromatic proteins in grapes (19).

Tandem duplication is another potent source of gene duplication. Tandemly duplicated genes represent ~15% of genes in angiosperm genomes (20); curiously, this proportion closely mirrors the 10 to 17% range of tandem duplicates found across animal genomes (21). Tandemly duplicated genes are probably subjected to different selection pressures than genes duplicated by polyploidy, on the basis of four lines of evidence: (i) First, tandem events tend to duplicate only one component of a genetic network, as opposed to entire networks. (ii) Second, tandemly duplicated genes are biased toward a different set of genes; tandem duplicates are overrepresented for membrane proteins and abiotic response genes (20). These genes tend to be at the end of biosynthetic pathways, suggesting that tandem duplicates are retained more readily if they do not perturb key branch points of networks. (iii) Third, differences between tandem

and polyploid duplicates extend to patterns of gene expression, because tandem duplicates diverge more rapidly in expression (22). (iv) Lastly, tandem duplication events are ongoing and common. It has been estimated conservatively that 1 out of ~700 *Arabidopsis thaliana* seeds contain a copy-number variant caused by unequal crossing over between tandem duplications (23).

Tandem duplication is common enough to occur on an ecological time scale and may thus

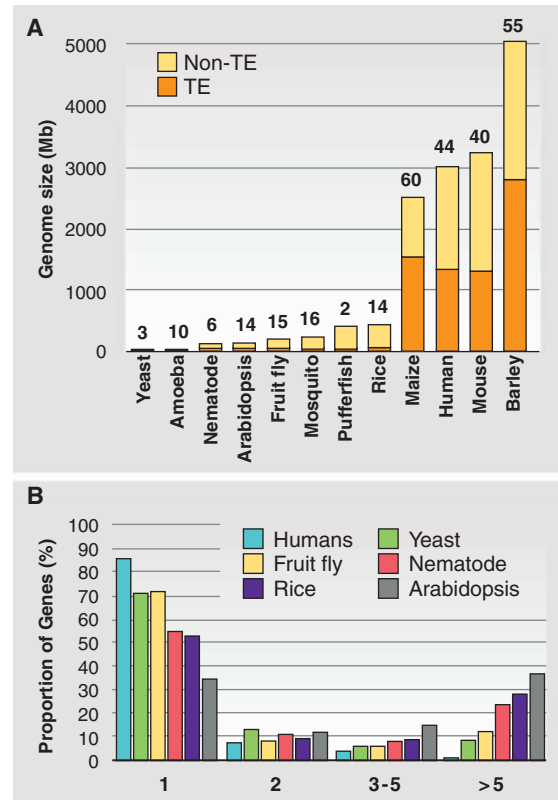
account for the 40% difference in genome size among some plant populations.

Our discussion underscores the need for genome-wide assessments of all types of genetic variation (including nucleotide polymorphism, copy-number variation, and TEs) at the population level. Such information is a necessary precursor for characterizing recent selection on plant genomes and also for understanding the mechanisms that contribute to genome-size variation.

Population genomic data can address the relative strengths of purifying, balancing, and directional selection; the genomic components that contribute to adaptation; and the identification of genes that have been targets of selection. These diversity assessments eventually need to include explicit multipopulation sampling so that diversity patterns can be evaluated to detect signatures of local adaptation. Thus far, the only genome-level polymorphism surveys in plants have targeted *A. thaliana* (28, 29), yielding insights about selection on coding regions and revealing unexpected trends (such as high levels of diversity in genes that mediate interactions with the biotic environment). Unfortunately, technological limitations inherent to these studies have prevented a comprehensive assessment of the relative frequency and amount of copy-number versus transposon polymorphism, and thus these important components of genome size and function remain poorly characterized.

Population genomic data provide information about recent selection, but inter-species comparisons may uncover selection manifested over longer time periods. Yet, there has been shockingly little comparative analysis of plant genomes, owing to the substantial evolutionary distances among the four angiosperm genome sequences published to date. This lack of analysis has highlighted the need for dense sequencing within

recently diverging clades. For example, the *A. lyrata* and sorghum genome sequences will provide fitting contrasts to those of *A. thaliana* and maize, respectively. These contrasts will yield information about the type and strength of selection on coding regions, molecular-evolutionary patterns that characterize species divergence, and basic dynamics of plant genome evolution. We do not yet know, for example, whether plant genomes contain large, conserved intergenic regions like those of animals and whether such intergenic regions constrain the lower limits of genome size. Similarly, we do not have a clear picture of the evolution of gene



**Fig. 2.** TE content and gene-family size for representative sets of eukaryote genomes. **(A)** Genomic TE content [from (8)]. Numbers above each bar represent the percentage of each genome made up of TEs. **(B)** Percent of the genome made up by gene families of varying sizes (single copy at left to greater than five copies at right) [from (31)].

be a particularly potent source of genetic innovation for local adaptation. Tandem duplications have been shown to mediate boron tolerance in barley (24), submersion tolerance in rice (25), and diversification of secondary metabolites in *A. thaliana* (26). However, there is not yet a great deal of information as to the extent of copy-number variation in plants, the role of copy-number variants in local adaptation, and the contribution of copy-number variants to genome-size variation among individuals. Careful characterization in humans indicates that up to 12% of the genome may vary in copy number (27), but even this impressive number is unlikely to

complement, particularly over modest (intra-familial) evolutionary distances. Comparative data will also facilitate a broader understanding of the dynamics of gene duplication and TE accumulation.

Nevertheless, additional comparative and population-genetic data alone will not yield a complete understanding of selection on plant genomes or on the processes that govern genome-size variation. There is first a pressing need for additional theoretical advances to provide a conceptual framework to interpret polymorphism data, especially in the context of demographic change in structured populations. Similarly, the theory of the population genetics of gene duplication is in its infancy, as is our understanding of whether standing genetic variation commonly contributes to adaptation. In addition, we need to better understand biological factors that affect the process of selection but are usually not included in molecular-evolutionary or population-genetic models; such factors include paramutation, methylation, epistasis, and gene conversion. Finally, there is always a need to complement inferences about selection with functional assays, particularly if the goal is to correctly identify the genetic variants that have been targeted by

selection. With the need for additional data and theoretical models, we clearly are only beginning to understand the complex interplay among phenotypic diversity, genome size, and natural selection.

## References and Notes

1. J. Greilhuber *et al.*, *Plant Biol. (Stuttgart)* **8**, 770 (2006).
2. T. R. Gregory *et al.*, *Nucleic Acids Res.* **35**, D332 (2007).
3. C. A. Knight, N. A. Molinari, D. A. Petrov, *Ann. Bot. (London)* **95**, 177 (2005).
4. A. L. Rayburn, H. J. Price, J. D. Smith, J. R. Gold, *Am. J. Bot.* **72**, 1610 (1985).
5. T. R. Meagher, C. Vassiliadis, *New Phytol.* **168**, 71 (2005).
6. A. L. Rayburn, J. W. Dudley, D. P. Biradar, *Plant Breed.* **112**, 318 (1994).
7. J. Ma, K. M. Devos, J. L. Bennetzen, *Genome Res.* **14**, 860 (2004).
8. M. G. Kidwell, *Genetica* **115**, 49 (2002).
9. K. Naito *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17620 (2006).
10. S. I. Wright, Q. H. Le, D. J. Schoen, T. E. Bureau, *Genetics* **158**, 1279 (2001).
11. D. E. Neafsey, J. P. Blumenstiel, D. L. Hartl, *Mol. Biol. Evol.* **21**, 2310 (2004).
12. M. Lynch, J. S. Conery, *Science* **302**, 1401 (2003).
13. Y. T. Aminetzach, J. M. Macpherson, D. A. Petrov, *Science* **309**, 764 (2005).

14. M. E. Hudson, D. R. Lisch, P. H. Quail, *Plant J.* **34**, 453 (2003).
15. J. D. Hollister, B. S. Gaut, *Mol. Biol. Evol.* **24**, 2515 (2007).
16. W. Wang *et al.*, *Plant Cell* **18**, 1791 (2006).
17. A. Torkamanzei, C. Moran, F. W. Nicholas, *Genetics* **131**, 73 (1992).
18. G. Blanc, K. H. Wolfe, *Plant Cell* **16**, 1679 (2004).
19. O. Jaillon *et al.*, *Nature* **449**, 463 (2007).
20. C. Rizzon, L. Ponger, B. S. Gaut, *PLoS Comput. Biol.* **2**, e115 (2006).
21. V. Shoja, L. Q. Zhang, *Mol. Biol. Evol.* **23**, 2134 (2006).
22. T. Casneuf, S. De Bodt, J. Raes, S. Maere, Y. Van de Peer, *Genome Biol.* **7**, R13 (2006).
23. B. S. Gaut, S. I. Wright, C. Rizzon, J. Dvorak, L. K. Anderson, *Nat. Rev. Genet.* **8**, 77 (2007).
24. T. Sutton *et al.*, *Science* **318**, 1446 (2007).
25. K. Xu *et al.*, *Nature* **442**, 705 (2006).
26. D. J. Kliebenstein, V. M. Lambrix, M. Reichelt, J. Gershenzon, T. Mitchell-Olds, *Plant Cell* **13**, 681 (2001).
27. R. Redon *et al.*, *Nature* **444**, 444 (2006).
28. R. M. Clark *et al.*, *Science* **317**, 338 (2007).
29. J. O. Borevitz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12057 (2007).
30. L. F. La Cour, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **285**, 61 (1978).
31. S. Lockton, B. S. Gaut, *Trends Genet.* **21**, 60 (2005).
32. We thank the Gaut lab for discussions. This work was funded by NSF grants to B.S.G.

10.1126/science.1153586

## PERSPECTIVE

# Synteny and Collinearity in Plant Genomes

Haibao Tang,<sup>1</sup> John E. Bowers,<sup>1</sup> Xiyin Wang,<sup>1</sup> Ray Ming,<sup>2</sup> Maqsudul Alam,<sup>3</sup> Andrew H. Paterson<sup>1\*</sup>

Correlated gene arrangements among taxa provide a valuable framework for inference of shared ancestry of genes and for the utilization of findings from model organisms to study less-well-understood systems. In angiosperms, comparisons of gene arrangements are complicated by recurring polyploidy and extensive genome rearrangement. New genome sequences and improved analytical approaches are clarifying angiosperm evolution and revealing patterns of differential gene loss after genome duplication and differential gene retention associated with evolution of some morphological complexity. Because of variability in DNA substitution rates among taxa and genes, deviation from collinearity might be a more reliable phylogenetic character.

Eukaryotic genomes differ in the degree to which genes remain on corresponding chromosomes (synteny) and in corresponding orders (collinearity) over time (1). For example, most eutherian (placental mammal) orders have incurred only moderate reshuffling of chromo-

somal segments since descent from common ancestors ~130 million years ago (2). Indeed, karyotype evolution along major vertebrate lineages appears to have been slow since an inferred whole-genome duplication occurred ~500 million years ago (3). Accordingly, accurate identification of orthologs across eutherian taxa is relatively routine, and deduction of synteny and collinearity is often straightforward with best-in-genome criteria (4), identifying one-to-one best matching chromosomal regions in pairwise genome comparisons.

Angiosperm (flowering plant) genomes fluctuate remarkably in size and arrangement even within close relatives, with recurring whole-

genome duplications occurring over the past ~200 million years accompanied by wholesale gene loss that has fractionated ancestral gene linkages across multiple chromosomes (5). Angiosperm genome sizes span more than 1000-fold (6), with much of the difference between some well-studied genomes in heterochromatin (7). Additionally, the reshuffling of short DNA segments by mobile elements nearly eliminates large-scale collinearity in heterochromatic regions (7).

Despite recurring whole-genome duplications, angiosperm chromosome numbers are more static than genome size, mostly within a range of less than 50-fold (6). Condensation of two chromosomes into one is known in many lineages; a particularly striking case involved the demonstration that  $n = 10$  (chromosome number) members of the *Sorghum* genus are ancestral to  $n = 5$  members of the genus (8). Indeed, *Sorghum bicolor* (sorghum) and *Zea mays* (maize) have the same chromosome number ( $n = 10$ ), although maize has been through a whole-genome duplication since their divergence (9), whereas the most recent duplication in sorghum is shared with all other cereals (10). The occurrence of several condensations may explain why single arms of several maize chromosomes (10 and 5) correspond to entire sorghum chromosomes (6 and 4) (11).

Fully sequenced genomes promise to improve deductions of correspondence, toward a unified framework for comparative evolutionary analysis.

<sup>1</sup>Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA. <sup>2</sup>Department of Plant Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA. <sup>3</sup>Advanced Studies in Genomics, Proteomics, and Bioinformatics Unit, University of Hawaii, Honolulu, HI 96822, USA.

\*To whom correspondence should be addressed. E-mail: paterson@uga.edu