EXTENDED PDF FORMAT
SPONSORED BY
usb
www.usbweb.com

Synteny and Collinearity in Plant Genomes
Haibao Tang, *et al.*
Science **320**, 486 (2008);
DOI: 10.1126/science.1153917

The following resources related to this article are available online at www.sciencemag.org (this information is current as of April 25, 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/320/5875/486

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/320/5875/486#related-content

This article **cites 24 articles**, 11 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/320/5875/486#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

Downloaded from www.sciencemag.org on April 25, 2008

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2008 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

complement, particularly over modest (intrafamilial) 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. Torkamanzehi, 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,[1] John E. Bowers,[1] Xiyin Wang,[1] Ray Ming,[2] Maqsudul Alam,[3] Andrew H. Paterson[1]*

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 chromosomal 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.

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA. [2]Department of Plant Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA. [3]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

Downloaded from www.sciencemag.org on April 25, 2008

In angiosperms, analysis of synteny and paleopolyploidy are inextricably intertwined because comparative genomics in angiosperm sequences require strategies to mitigate the effects of genome duplication and fractionation. For example, *Arabidopsis thaliana* (thale cress) has undergone three paleo-polyploidies, including two doublings (*5*) and one tripling (*12*), resulting in ~12 copies of its ancestral chromosome set in a ~160-Mb genome. Further complicating the comparison of *A. thaliana* to other angiosperms are an additional 9 to 10 chromosomal rearrangements in the past few million years since its divergence from *A. lyrata* (rock cress) and *Capsella rubella* (pink shepherd's purse), including condensation of six chromosomes into three, bringing the chromosome number from $n = 8$ to $n = 5$ (*13*).

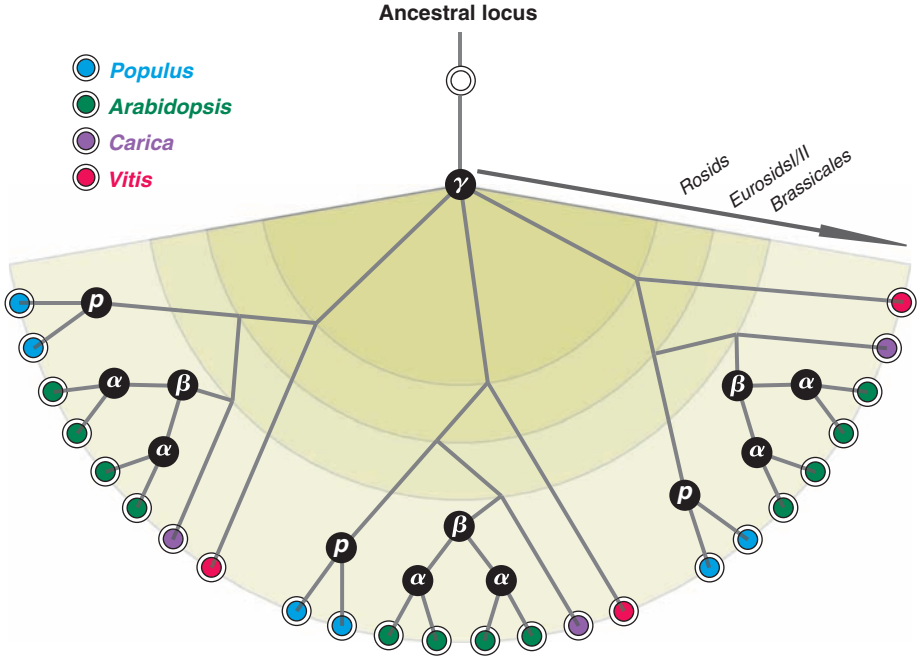

Fig. 1. Idealized gene tree that contains multiple orthologs and paralogs in *Populus*, *Arabidopsis*, *Carica*, and *Vitis*. For illustration purpose, this has assumed equal evolutionary rates along all branches and no gene loss following polyploidy. The polyploidy events are represented as black circles and labeled α and β within the *Arabidopsis* lineage (*5*), salicoid duplication *p* in *Populus* (*15*), and γ, which is shared by all four species (*12*, *14*).

Other eudicot genomes show less-complicated genome architectures than *Arabidopsis*. Although still controversial, the two most recent paleopolyploidies affecting *Arabidopsis* [α and β, following the usage in (*5*)] now appear to have occurred within the crucifer lineage (*12*, *14*). *Populus trichocarpa* (poplar) underwent a duplication specific to its own salicoid lineage (*15*) and shares only one of the three paleo-polyploidies (γ) affecting *Arabidopsis*. *Vitis vinifera* (grape) (*12*) and *Carica papaya* (papaya) (*14*), the latter within the same taxonomic order (Brassicales) as *Arabidopsis*, each have only γ and no subsequent polyploidies (Fig. 1). Indeed, the absence of the β event in *Carica* (*14*) argues against an alternative interpretation on the basis of an analysis of a second *Vitis* genome (*16*), which suggested that β occurred in a common ancestor of *Arabidopsis-Populus*.

Synteny can be identified through the clustering of neighboring matching gene pairs; however, differences in gene density and tandem gene arrays among species may cause statistical artifacts. Collinearity, a more specific form of synteny, requires common gene order. Collinearity and synteny have traditionally been identified by looking for one-to-one (pairwise) conservation between species. To take better advantage of new genomic resources as they become available, multiway collinearity analyses are needed, with progressive alignments accompanied by statistical evaluation and iterative refinement (*4*). In angiosperms, such multiple alignments offer the further advantage of helping to unravel the consequences of genome duplications.

One partial solution for inferring ancestral gene orders in angiosperms has been a bottom-up approach, in which the most recently duplicated segments are interleaved to generate hypothetical intermediates that are further recursively merged (*5*). However, this approach requires an additional cycle of deductions for each duplication event and compounds any errors. An alternative top-down approach requires only one cycle of deduction by simultaneously searching for and aligning all structurally similar segments across multiple genomes and subgenomes. The top-down approach should be more sensitive because it can incorporate transitive homology (*17*), in which segments A and B have undergone reciprocal gene loss and no longer show correspondence to each other but both correspond with a third segment, C. Relationships among such degenerated duplicated regions, easily missed by a bottom-up approach, can often be resolved by comparison to another genome that does not have the duplication or that underwent independent gene loss. Such comparisons have clarified synteny among yeast species (*18*).

Top-down analyses show a high degree of collinearity between *Arabidopsis*, *Carica*, and *Populus* (*14*). For example, we identified three branches each containing orthologous segments from up to four *Arabidopsis*, one *Carica*, and two *Populus* genomic region(s), suggesting paleohexaploidy in a common ancestor of these species (Fig. 2A). Applying these methods to the *Vitis* (grape) genome validated the reconstructed order and inferred triplicated structure of a common *Arabidopsis-Carica-Populus* ancestor. *Vitis* is a eudicot outside of the two eurosid clades that contain *Arabidopsis-Carica* (eurosids II) and *Populus* (eurosids I) (*19*), therefore providing an independent lineage suitable to test the gene order alignments. Paleo-hexaploidy (triplication) has also been suggested over 94.5% of the *Vitis* genome (*12*). When the *Arabidopsis-Carica-Populus* consensus is aligned to *Vitis*, the two independently inferred triplication patterns correspond closely (Fig. 2B). Thus, top-down gene order alignment revealed genome triplication that eluded prior detection in *Arabidopsis* (*5*) and *Populus* (*15*) and also supported the conclusion that the triplication occurred in a common ancestor of *Vitis*, *Arabidopsis*, *Carica*, and *Populus* (*12*).

The emerging unified framework for comparative evolutionary analysis of angiosperm genes and genomes will improve in power and precision as more genomes are sequenced. However, the current framework remains bipolar because we can identify extensive synteny and collinearity within core eudicots and grasses, respectively, but much less between the two groups because of longer evolutionary distance and more genome rearrangements. Collinear orthologs between rice (*Oryza sativa*) and the four core eudicots account for only ~15% of *Oryza* genes distributed over about half of the genome. The longest *Oryza-Arabidopsis* collinear segment contains 23 orthologous gene pairs but is improved twofold, to 47, by incorporating *Vitis*. Additional monocot sequences from noncereal genomes such as *Musa acuminata* (banana) or *Ananas comosus* (pineapple), along with sequences of basal eudicots such as *Eschscholzia californica* or *Papaver somnifera* (California or opium poppy) and *Aquilegia formosa* (columbine), and basal angiosperms such as *Amborella trichopoda* (no common name), may further

Downloaded from www.sciencemag.org on April 25, 2008

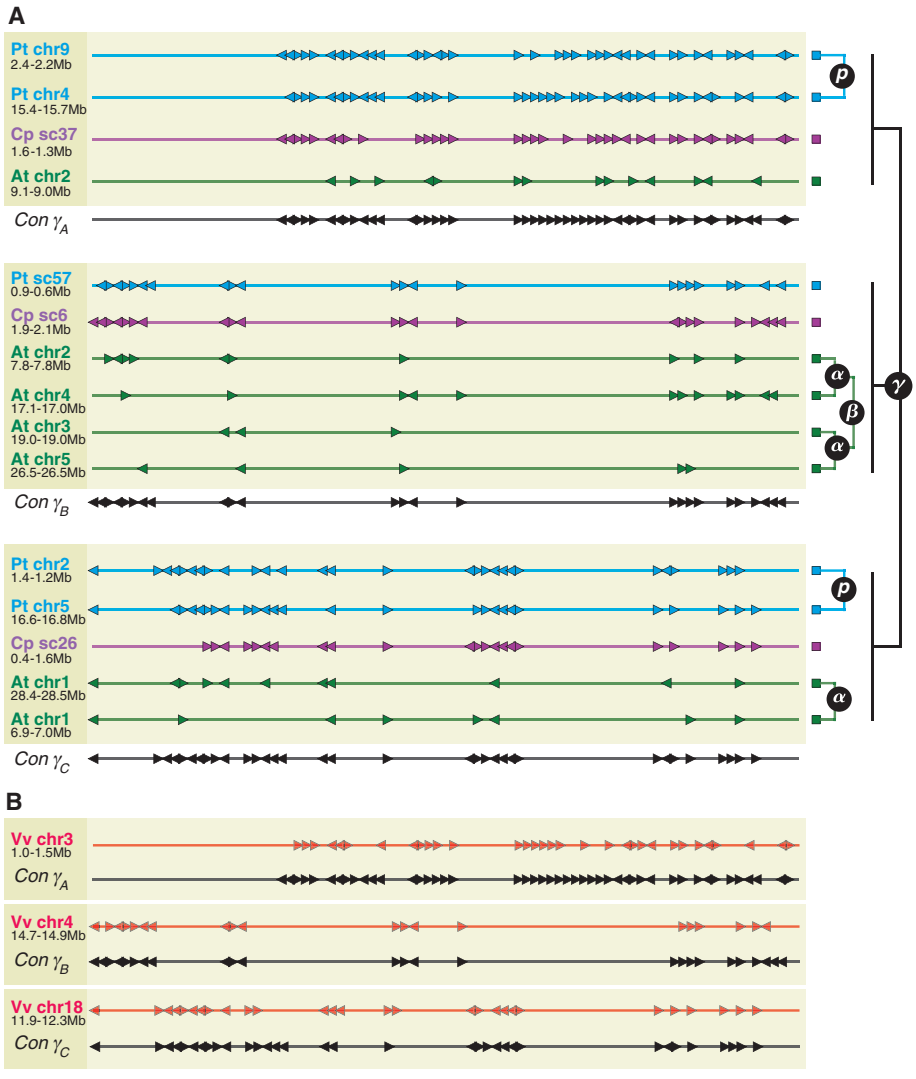

Fig. 2. Typical view of multiple collinear regions among several eudicot genomes. Triangles represent individual genes and their transcriptional orientations. Genes with no syntenic matches to the selected regions are not plotted. (**A**) Alignment among *Arabidopsis* (green), *Carica* (magenta), and *Populus* (blue) chromosomal regions. The whole alignment reveals four distinct duplications, illustrated in Fig. 1. The regions are grouped into three consensus γ-subgenomes (Con γ_A, γ_B, γ_C) on the basis of parsimony. Aligned genes within each γ subgenome are merged into an inferred order by consensus. (**B**) The inferred γ partitions are validated with the *Vitis* genome (red) because each γ subgenome clustered in (A) has only one closely matching *Vitis* chromosomal region.

improve detection of collinearity and synteny across major angiosperm clades.

Pan-angiosperm genome comparisons show correlated patterns of gene retention and loss in paleo-polyploid lineages. Alignments of multiple descendant chromosomes after polyploidy events reveal cases in which ancestral genes were deletion-resistant, consistently being preserved in syntenic subgenomes (*20*). Such preferential conservation of genes from particular families such as MADS-box genes (*21*) and other transcription factors may contribute to increasing morphological complexity (*22*). The opposite case is that of gene functional groups for which members have been consistently restored to one copy after multiple polyploidy cycles, suggesting that there are advantages in having only single copies of these genes (*20*).

Because of variability in DNA substitution rates among plants, deviation from collinearity might be a more reliable phylogenetic character. DNA substitution rates can be highly variable among seed plant lineages, with extreme cases showing 100-fold variation within the same genus on the basis of a study of mitochondrial genes (*23*). Analysis of rare changes (when compared to DNA substitutions) in genomic structure—such as specific rearrangements of gene order, insertions, or deletions—provides an informative and robust way to resolve relationships among many lineages (*24*). In retrospect, early inferences on polyploidy in angiosperms and vertebrates were initially confused by gene phylogenies but later resolved with synteny (*12*, *25*).

Improved synteny and collinearity alignments emerging from top-down approaches applied to multiple genomes and subgenomes are a potential foundation for reconstruction of the ancestral state(s) of angiosperm genomes. Consensus gene orders within syntenic blocks can be approximated on the basis of top-down alignments. Ordering among the syntenic blocks themselves on the macrolevel is more difficult; however, several combinatorial algorithms exist to reconstruct ancestral genomes under a most-parsimonious rearrangement scenario (*26*). The resulting orders would reveal not only shared but also divergent genes inserted into novel locations, underlining lineage-specific changes. Additional genome sequences will improve power to resolve gene orders at the microlevel and also contribute to identifying functionally important DNA, such as the evolutionarily constrained elements among 28 vertebrate genomes (*4*).

References and Notes

1. A. Coghlan, E. E. Eichler, S. G. Oliver, A. H. Paterson, L. Stein, *Trends Genet.* **21**, 673 (2005).
2. M. A. Ferguson-Smith, V. Trifonov, *Nat. Rev.* **8**, 950 (2007).
3. Y. Nakatani, H. Takeda, Y. Kohara, S. Morishita, *Genome Res.* **17**, 1254 (2007).
4. W. Miller *et al.*, *Genome Res.* **17**, 1797 (2007).
5. J. E. Bowers, B. A. Chapman, J. Rong, A. H. Paterson, *Nature* **422**, 433 (2003).
6. M. D. Bennett, J. B. Smith, *Philos. Trans. R. Soc. Ser. B* **334**, 309 (1991).
7. J. E. Bowers *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13206 (2005).
8. R. Spangler, B. Zaitchik, E. Russo, E. Kellogg, *Syst. Bot.* **24**, 267 (1999).
9. Z. Swigonova *et al.*, *Genome Res.* **14**, 1916 (2004).
10. A. H. Paterson, J. E. Bowers, B. A. Chapman, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9903 (2004).
11. J. E. Bowers *et al.*, *Genetics* **165**, 367 (2003).
12. O. Jaillon *et al.*, *Nature* **449**, 463 (2007).
13. K. Yogeeswaran *et al.*, *Genome Res.* **15**, 505 (2005).
14. R. Ming *et al.*, *Nature*, in press.
15. G. A. Tuskan *et al.*, *Science* **313**, 1596 (2006).
16. R. Velasco *et al.*, *PLoS One* **2**, e1326 (2007).
17. Y. Van de Peer, *Nat. Rev. Genet.* **5**, 752 (2004).
18. M. Kellis, B. W. Birren, E. S. Lander, *Nature* **428**, 617 (2004).
19. D. E. Soltis, P. S. Soltis, P. K. Endress, M. W. Chase, *Phylogeny and Evolution of Angiosperms* (Sinauer, Sunderland, MA, 2005).
20. A. H. Paterson *et al.*, *Trends Genet.* **22**, 597 (2006).
21. J. Nam, C. W. dePamphilis, H. Ma, M. Nei, *Mol. Biol. Evol.* **20**, 1435 (2003).
22. M. Freeling, B. C. Thomas, *Genome Res.* **16**, 805 (2006).
23. J. P. Mower, P. Touzet, J. S. Gummow, L. F. Delph, J. D. Palmer, *BMC Evol. Biol.* **7**, 135 (2007).
24. A. Rokas, P. W. Holland, *Trends Ecol. Evol.* **15**, 454 (2000).
25. P. Dehal, J. L. Boore, *PLoS Biol.* **3**, e314 (2005).
26. E. E. Eichler, D. Sankoff, *Science* **301**, 793 (2003).

27. Funded by NSF MCB-0450260 to A.H.P. and J.E.B., DBI-0421803 to R.M. and A.H.P., the U. Hawaii to M.A., and U.S. Department of Defense W81XWH0520013 to M.A.

10.1126/science.1153917

Downloaded from www.sciencemag.org on April 25, 2008