

confirmed by both electron-polymerase chain reaction (e-PCR) and fluorescent in situ hybridization (FISH) analyses (<http://ncbi.nlm.nih.gov/genome/clone>). In addition, PTB-053J22 is fully included in another independently sequenced chimpanzee BAC clone (AC007214, RP43-135M19). Thus, it is highly likely that PTB-053J22, which we detected through this study, contains one of the breakpoints corresponding to the human (or vice versa in chimpanzee) chromosomal inversion between 12p12 and 12q15. In addition, this region in human chromosome 12 or in chimpanzee chromosome 10 is known to be inverted in gorilla and chimpanzee as opposed to human and orangutan (8, 18). We found several genes around the PTB-053J22 BESs in the following order, SCL21A14 (solute carrier, organic anion transporter, family 21, member 14), <36 kb>, PTB-053J22-F, <5 kb>, SLC21A8 (solute carrier, organic anion transporter, family 21, member 8, <cen>, DYRK2 (dual-specificity tyrosine-Y-phosphorylation regulated kinase 2), <330 kb>, PTB-053J22-R, <168 kb>, and IFNG (interferon- γ), based on the annotations on the corresponding NT contigs. The effect of the inversion on these genes should be the target of future studies.

To independently test our mapping procedure, we selected 15 chimpanzee BAC clones mapped to human chromosomes 1 to 8 by the BES alignment procedure (13) and subjected them to FISH analysis with both human and chimpanzee M-phase cell spreads (Fig. 3). As shown, 13 clones showed single locus signals at the corresponding positions on both human and chimpanzee chromosomes, and two clones, RP43-50L24 and RP43-60K09, showed similar signals at two loci on the human and chimpanzee chromosomes, suggesting that the mapping procedure we used in this study is reliable. We believe that the whole genome chimpanzee/human comparative map built here by the BES alignment procedure is reasonably accurate and useful for future studies. Recent development of the human-mouse comparative map (19, 20) also supports our approach.

Users of this map should still be careful in applying the information because the possibility remains that assignment of particular clones in the NT contig is incorrect or that inter- or intrachromosomally duplicated regions may be included within an insert. However, the quality of the map, and thus its usefulness, should increasingly improve as the finishing of the human genome sequence proceeds.

References and Notes

1. International Human Genome Sequencing consortium, *Nature* **409**, 860 (2001).
2. J. C. Venter et al., *Science* **291**, 1304 (2001).
3. M.-C. King, A. C. Wilson, *Science* **188**, 107 (1975).

4. A. Jauch et al., *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8611 (1992).
5. B. Dutrillaux, *Hum. Genet.* **48**, 251 (1979).
6. W. Burrows, O. A. Ryder, *Nature* **385**, 125 (1997).
7. M. Goodman et al., *Mol. Phylogenet. Evol.* **9**, 585 (1998).
8. E. Nickerson, D. L. Nelson, *Genomics* **50**, 368 (1998).
9. J. G. Hacia et al., *Nature Genet.* **18**, 155 (1998).
10. H. Kaessmann, V. Wiebe, S. Paabo, *Science* **286**, 1159 (1999).
11. P. Gagneux, A. Varki, *Mol. Phylogenet. Genet.* **18**, 2 (2001).
12. S. Paabo, *Science* **291**, 1219 (2001).
13. Methods to construct the BAC end sequences (BESs) library and the comparative map can be seen at <http://hgp.gsc.riken.go.jp> or *Science* Online (14). All sequence reads were submitted to the DNA Data Bank of Japan (DDBJ) under accession numbers AG029037 to AG186569 and AG186783 to AG187837.
14. The supplemental data are available at hgp.gsc.riken.go.jp or www.sciencemag.org/cgi/content/full/295/5552/131/DC1.
15. The probability equation, $N = \ln(1 - P)/\ln(1 - 1/n)$, where N is the number of clones, in this case 24,580, and n is the total size (2.7 Gb in this case) divided by the size of fractions. The estimated insert size of the BAC libraries is about 130 to 150 kb. The calculated P value is 0.694 (130-kb insert) or 0.745 (150-kb insert).
16. M. Hattori et al., *Nature* **405**, 311 (2000).
17. In total, 752 primer pairs were designed from the human chromosome 21 sequence at distances of about 50 kb (sequences of the primer pairs can be seen at *Science* Online) (14). For the reactions successful only for human DNA, we expanded the test to include two additional genomic DNA preparations from different individuals of chimpanzee (male and female) and five other primates, including gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), Old World monkey (*Macaca fascicularis*), New World monkey (*Ateles geoffroyi*), and prosimian (*Lemur catta*).
18. International System for Human Cytogenetic Nomenclature (1985) (ISCN1985) (Karger, Basel, Switzerland, 1985), pp. 95–109.
19. D. Butler, *Nature* **413**, 444 (2001).
20. See <http://mouse.ensembl.org>.
21. We thank all the members and associates (individual names can be seen at our Web site) of the Human Genome Research Group, RIKEN Genomic Sciences Center, for their enthusiastic support and stimulating discussions. Special thanks to RIKEN GSC and the National Institute of Genetics for supporting the international workshop on chimpanzee genomics, held in Tokyo, March 2001. This work was supported in part by special grants from the Ministry of Education, Culture, Sports, Science and Technology for Genome Research.

8 August 2001; accepted 13 November 2001

Nucleotide Variation Along the *Drosophila melanogaster* Fourth Chromosome

Wen Wang,¹ Kevin Thornton,² Andrew Berry,³ Manyuan Long^{1,2*}

The *Drosophila melanogaster* fourth chromosome, believed to be nonrecombining and invariable, is a classic example of the effect of natural selection in eliminating genetic variation in linked loci. However, in a chromosome-wide assay of nucleotide variation in natural populations, we have observed a high level of polymorphism in a ~200-kilobase region and marked levels of polymorphism in several other fragments interspersed with regions of little variation, suggesting different evolutionary histories in different chromosomal domains. Statistical tests of neutral evolution showed that a few haplotypes predominate in the 200-kilobase polymorphic region. Finally, contrary to the expectation of no recombination, we identified six recombination events within the chromosome. Thus, positive Darwinian selection and recombination have affected the evolution of this chromosome.

Detecting evolutionary forces that shape the structure of genetic variation at the genomic level often relies on understanding the effects of natural selection on nearby linked loci (1–4), for which the fourth chromosome of *Drosophila melanogaster* is a classical model system (5–7). It has been thought to undergo no meiotic recombination except under certain experimental conditions (e.g., the interchromosomal effect introduced for the purpose of mapping)

(8, 9). Two genetic models—"selective sweep" by the hitchhiking effect (1, 10), in which an advantageous allele is selected for and fixed in species rapidly, and the background selection model, in which deleterious mutations are selected against (2, 10)—predicted a lack of variation throughout the chromosome, which has been supported by limited data (6, 11). We have reexamined the level of variation and recombination in the fourth chromosome using a chromosome-wide assay of nucleotide variation.

We first investigated within-species nucleotide variation in two adjacent regions from the 102F cytological position of the fourth chromosome of *D. melanogaster*: (i) 4257 base pairs (bp) of the *CG11091* locus, and (ii) 847 bp of an intron of the *toy* gene. *CG11091* and *toy* are separated by ~10 kb. We directly

¹Department of Ecology and Evolution, ²Committee on Genetics, University of Chicago, 1101 East 57 Street, Chicago, IL 60637, USA. ³Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: mlong@midway.uchicago.edu

Both the worldwide and population samples show high levels of nucleotide variation in the *CG11091-toy* region. The entire region of 5104 nucleotides (nt) contains 47 segregating sites for the worldwide sample (nucleotide diversity $\pi = 0.0028$) and 32 segregating sites for the IS population ($\pi = 0.024$). These segregating sites include six insertion/deletion (indel) sites. The π value for the *toy* gene is as high as 0.0049 for the worldwide sample and 0.0043 for the IS population. This level of variation is an order of magnitude greater than those previously observed in *ci* genes in the fourth chromosomes of *D. melanogaster* and *Drosophila simulans* (6) and in *Drosophila sechellia* and *Drosophila mauritiana* (11), for which the values range from 0 to 0.0003. The probability distribution of segregating sites (12) in these IS and worldwide samples revealed that the observed levels of variation are significantly higher than the nucleotide diversities of *ci* (0.0002) (6, 11) and other loci in regions of low recombination (0.0005) (13) ($P < 0.001$). Indeed, the observed levels of variation are in the range typically seen for autosomal genes in regions of normal recombination (13).

along the fourth chromosome is partitioned into two distinct sets of haplotypes with unequal frequencies. Hereafter, we refer to the high- and low-frequency haplotype groups as the major and minor haplotypes, respectively. The frequency of the major haplotype is similar in both the IS and worldwide data sets (8/11 and 14/23, respectively). Furthermore, these two haplotypes appear in all local populations tested (Fig. 1) (14), suggesting that the haplotype distribution in the IS population is typical worldwide. These results raise new questions about the evolutionary forces shaping variation and recombination on the fourth chromosome, prompting further analysis.

Using the 11 randomly sampled alleles from the IS population (Fig. 1), we carried out three different statistical tests of the null hypothesis of neutrality (15), assuming randomly generated haplotypes: haplotype partition test (HP test) (16), haplotype number test (K test) (17), and haplotype diversity test (H test) (17). We also calculated Tajima's D value (18), a measure of skewness in the frequency spectrum of polymorphic sites. For all three tests, probability values were estimated by Monte Carlo simulation (19). Because the sequenced region of *toy* is too short for powerful statistical testing and its pattern and levels of polymorphism are similar to those of *CG11091*, we pooled data from both gene regions, comprising 32 segregating sites and five haplotypes with a haplotype diversity of 0.618.

An alternative interpretation is that the departure from the null hypothesis of neutrality is a consequence of ancient admixture of two differentiated populations, consistent with a demographic cause (20). However, the worldwide survey shows that all local populations contain both haplotypes, providing no evidence for population differentiation in terms of variation along the fourth chromosome. Furthermore, such a model predicts similar dimorphisms elsewhere in the genome; this has not been found in the many population genetic studies of *D. melanogaster* variation. Thus, it is more parsimonious to interpret the observed dimorphism as a consequence of balancing selection.

We also examined whether the haplotype structure was associated with a chromosomal inversion that may be under balancing selection, resulting in the observed pattern. We used fluorescence in situ hybridization (FISH) of

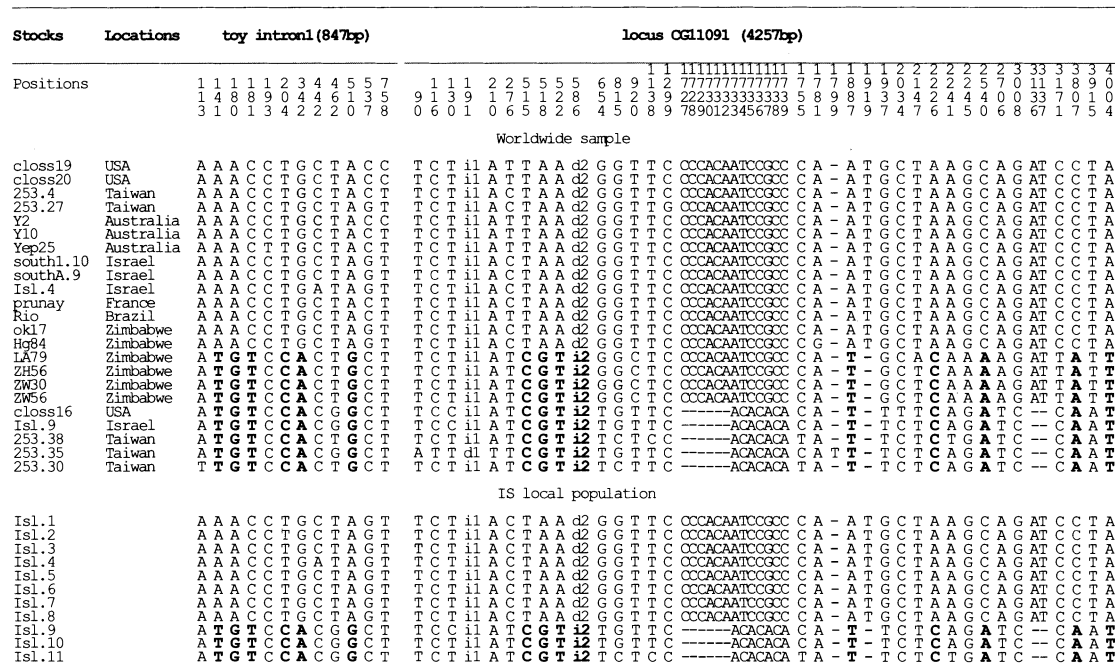


Fig. 1. Nucleotide variations in the *CG11091-toy* region of the *D. melanogaster* fourth chromosome. The sites defining the two haplotype groups are shown in bold type in the low-frequency haplotype group. The numbers for the positions, e.g., 113 and 90, indicate the positions of polymorphic sites in *toy* (site 113 to 758) and *CG11091* (site 90 to 4004).

i1/d1 and i2/d2 are two insertion-deletion (indel) polymorphisms: i1 indicates presence of the sequence ATTTTACAAA and d1, absence of this sequence; i2 indicates presence of the sequence GGTGTATCATTT-GCTTTC and d2, absence of this sequence. Other indel polymorphisms are shown directly; dashes indicate absence of nucleotides.

polytene chromosomes (21) to ascertain the gene order within the haplotype region; this did not reveal any inversion. An alternative scenario of positive selection would involve the major haplotype undergoing a selective sweep, being driven by selection toward fixation. However, that the two haplotypes are polymorphic in all locations argues against this possibility.

Because our results contrast with those from previous studies of *ci* in *Drosophila*, we sequenced 10 *ci* genes from our worldwide sample and again found no polymorphism (Fig. 2). The lack of variation at *ci* implies different evolutionary histories in different chromosomal regions and the existence of recombination between *ci* and *CG11091-toy*. We investigated these conjectures by examining patterns of linkage disequilibrium throughout the euchromatic region of the chromosome by sequencing an additional 15 gene regions (22) (Fig. 2). We also sequenced seven related genes in *D. simulans*, the sibling of *D. melanogaster*, to further investigate the role of selection and to determine the ancestral haplotype.

First, we found that the chromosome could be divided into three discrete domains: (i) the 200-kb dimorphic domain containing the *CG11091-toy* region with its characteristic two-haplotype organization of variation; (ii) a polymorphic proximal domain in which no such haplotype organization is seen; and (iii) a domain, distal to the centromere, where levels of variation are low. The first domain has the highest level of variation in comparison with the two other domains, although its average silent nucleotide diversity, 0.0021, is lower than the average nucleotide diversity in the genome (13). The latter two domains, both proximal and

distal to the dimorphic domain, show no such dimorphism, and varying levels of polymorphism. Although many genes in these two domains show levels of polymorphism characteristic of regions of reduced recombination (13), 3 out of 11 gene regions in these areas show relatively high levels of nucleotide diversity (0.0012 to 0.0019). The boundaries between these domains can be narrowed to two short regions of 15 kb (boundary 1 between the regions *CG11153* and *B*) and 7 kb (boundary 2 between *toy* and *plexA*).

A statistical test for heterogeneity of variation among gene regions, based on the χ^2 - Kreitman-Hudson test statistic (23), shows, for all 18 gene regions, significant heterogeneity ($P \ll 0.0001$). This further indicates that the fourth chromosome is not evolving as a single unit; different regions appear to have different evolutionary histories.

All but one of the five gene regions in the dimorphic domain in the additional survey show the same haplotype structure as the *CG11091-toy* region; locus *A*, however, shows no variation. This domain thus displays linkage disequilibrium over some 200 kb. The major haplotype is less diverse than the minor haplotype: The major haplotype cluster contains 9 segregating sites (one indel site included), whereas the minor haplotype cluster contains 30 such sites (2 indel sites included) (Fig. 2). In the *CG11091-toy* region, the worldwide sample reveals that for the within-minor haplotype group $\pi = 0.0015$, and for the within-major haplotype group $\pi = 0.0004$. Consistent with these observations, the *D. simulans* outgroup sequences in gene regions *CG11152* and *CG11091* reveal that the minor haplotype is ancestral.

By pooling all 16 loci that contain polymorphisms, we have estimated a minimum number of six recombination events (R_m) by the four-gamete method (Fig. 2) (24). However, because we have not obtained a continuous sequence along the entire chromosome, the true R_m for the fourth chromosome is in all probability larger. Thus, to calculate an upper bound on R_m , we identified one event in *CG11093* from 20,010 nt that we sequenced, yielding a R_m density of $1/20,010 = 0.05/\text{kb}$ in 10 chromosomes. To calculate a lower bound, we assume a R_m density of six events/1156-kb nucleotides (the chromosomal euchromatin length) = $0.0052/\text{kb}$ in these chromosomes. Although these results cannot be directly compared with the experimental recombination estimates, it is informative to compare them with the estimates from other population genetic data. The *Adh* gene in a moderate-recombination region has a R_m density of 1.84 events per kilobase in 11 alleles (24). Thus, qualitatively, rates of recombination on the fourth chromosome are 37- to 354-fold lower than those on normal autosomes. The amount of recombination observed here is low, consistent with genetic analysis of the chromosome (8). In contrast to previous predictions (1-3, 6), however, such a low rate has a considerable effect on the structure of genetic variation on the chromosome.

Recombination caused by crossovers at each end of the dimorphic domain may account for the different evolutionary histories of the three chromosome domains as described above. The genome sequence (25) reveals that both putative boundary regions contain many repetitive sequences that may facilitate genetic recombination.

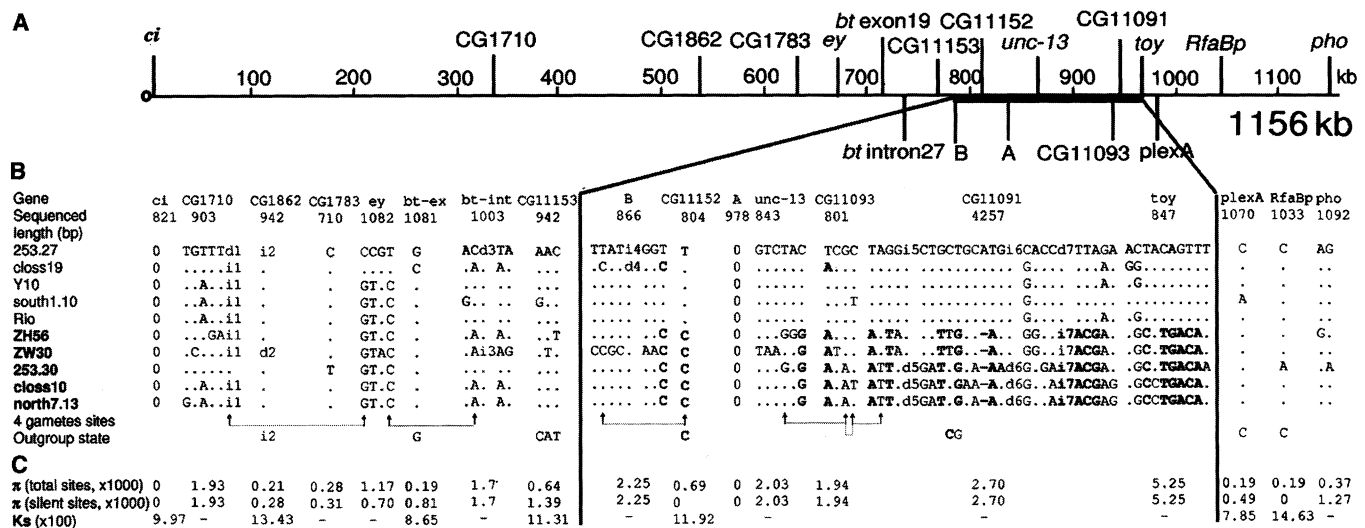


Fig. 2. Distribution of variation on the *D. melanogaster* fourth chromosome. (A) The surveyed 18 gene regions along the *D. melanogaster* chromosome in the map (25) with the gene order between *CG11153* and *pho* corrected on the basis of our FISH experiments. (B) Segregating sites in the nonrandomly sampled 10 chromosomes with dimorphism shown in bold type. Arrows indicate the minimum number of recombination (R_m) sites identified by the

four-gamete method (24). *Drosophila simulans* is the outgroup. i/d denotes indel polymorphism of various lengths indicating presence (i) or absence of nucleotide sequences (d). Seven indel polymorphic sites are shown: i1/d1 = 16 bp, i2/d2 = 2 bp, i3/d3 = ~1 kb, i4/d4 = 13 bp, i5/d5 = 2 bp, i6/d6 = 11 bp, i7/d7 = 18 bp. (C) K_s : synonymous divergence per nucleotide site between *D. melanogaster* and *D. simulans*.

Given that recombination does occur on the fourth chromosome, the maintenance of the huge dimorphic domain is anomalous—we would expect it to be eroded by recombination. However, it seems plausible to suppose that the dimorphism is the joint product of balancing selection on a locus within the region, and a low rate of recombination such that variation linked to one balanced allele is seldom, if ever, recombined into association with the other allele.

The significantly reduced variation outside the dimorphic domain could be due to either a reduced mutation rate, hitchhiking with positive Darwinian selection, or background selection. The first hypothesis, which predicts that low divergence between species will correspond to low variation within species, was not supported by the observed typical level of silent site substitutions, K_s ($K_s = 0.0785 \sim 0.1463$) (Fig. 2) (3, 6). For the second and third hypotheses, a Tajima's D test on pooled data from all seven gene regions in the centromere-proximal nondimorphic domain shows no significant bias in the polymorphism spectrum ($D = -0.9745$, $P = 0.1739$) and thus does not support a recent selective sweep over this long region (26). This leaves the possibility that other forms of selection—e.g., background selection or directional selection in local regions delineated by recombination—may play a role. Even if selective sweep does occur in some local regions, the low recombination rate would render it a slow process and make it unlikely to be global.

Previous studies, both theoretical and empirical, had concluded that the fourth chromosome lacks variation. However, we have found that it not only harbors high levels of nucleotide variation throughout the chromosome, but also has a unique dimorphism that extends across a long chromosome domain, suggesting the importance of positive Darwinian selection (balancing selection) in the evolution of this chromosome. These results may be viewed as empirical support for Dobzhansky's "coadapted gene complex" idea (27), with each haplotype representing a distinct complex. The evolution of such a complex—if it is to occur at all—is most likely to occur in regions of low recombination like the one in question. These results provide a starting point for reassessing the genetic and evolutionary forces that affect both

this chromosome in particular, and low recombination regions in general.

References and Notes

1. J. Maynard-Smith, J. Haigh, *Genet. Res.* **23**, 23 (1976).
2. B. Charlesworth, M. T. Morgan, D. Charlesworth, *Genetics* **134**, 1289 (1993).
3. D. J. Begun, C. F. Aquadro, *Nature* **356**, 519 (1992).
4. D. Nurminsky, D. D. Aguiar, C. D. Bustamante, D. L. Hartl, *Science* **291**, 128 (2001).
5. B. Charlesworth, *Nature* **356**, 475 (1992).
6. A. J. Berry, J. W. Ajioka, M. Kreitman, *Genetics* **129**, 1111 (1991).
7. S. Freeman, J. C. Herron, *Evolutionary Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1998).
8. B. Hochman, in *The Genetics and Biology of Drosophila*, M. Ashburner, E. Novitski, Eds (Academic Press, New York, 1976), vol. 1b, pp. 903–928.
9. M. Ashburner, *Drosophila: A laboratory handbook* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).
10. B. Charlesworth, *Genet. Res.* **68**, 131 (1996).
11. H. Hilton, R. M. Kliman, J. Hey, *Evolution* **48**, 1900 (1994).
12. R. R. Hudson, *Oxford Survey Evol. Biol.* **7**, 1 (1990).
13. J. R. Powell, *Progress and Prospects in Evolutionary Biology, the Drosophila Model* (Oxford Univ. Press, New York, 1997).
14. We also assayed 11 additional isofemale lines from South America, five additional lines from Indiana, three additional lines from Australia, and two additional lines from France by PCR using haplotype-specific primers. Both the major and the minor haplotypes were detected in all these populations.
15. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
16. R. R. Hudson, K. Bailey, D. Skarecky, J. Kwiatkowski, F. J. Ayala, *Genetics* **136**, 1329 (1994).
17. F. Depaulis, M. Veuille, *Mol. Biol. Evol.* **15**, 1788 (1998).
18. F. Tajima, *Genetics* **123**, 585 (1989).
19. We performed simulations using a modification of the method described by Hudson (12), under a conservative assumption of no recombination. Rather than performing simulations using θ as a parameter, we instead randomly generated genealogies and then placed the observed number of polymorphic sites onto them (16). Here, $\theta = 4N\mu$, where N and μ are effective population size and neutral mutation rate, respectively. All probability values were estimated as one-tailed probabilities from 10,000 simulations [$P(X \leq X_{obs})$].
20. P. Andolfatto, M. Przeworski, *Genetics* **156**, 257 (2000).
21. W. Wang, J. Zhang, C. Alvarez, A. Llopart, M. Long, *Mol. Biol. Evol.* **17**, 1294 (2000).
22. We chose 10 individuals representing five major haplotype lines and five minor ones for sequencing an additional 15 gene regions on the fourth chromosome by PCR. The same male DNA sample from each line was used for PCR amplification of all 18 gene regions surveyed in this study. All newly created sequences for Figs. 1 and 2 have been deposited in GenBank (accession numbers AF433680 to AF433874 and AF461436 to AF461455).
23. M. Kreitman, R. R. Hudson, *Genetics* **127**, 565 (1991).
24. R. R. Hudson, N. L. Kaplan, *Genetics* **111**, 147 (1985).
25. M. D. Adams et al., *Science* **287**, 2185 (2000).
26. J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, W. Stephan, *Genetics* **140**, 783 (1995).
27. Th. Dobzhansky, *Genetics* **28**, 162 (1943).
28. We thank M.-L. Wu, E. Nevo, W. Ballard, and P. Gilbert for *Drosophila* strains and R. R. Hudson, M. Kreitman, J. Spofford, C.-I. Wu, B. Charlesworth, R. C. Lewontin, C. H. Langley, E. Stahl, and members of the Long lab for helpful discussions. Supported in part by grants from NSF, a Packard Fellowship for Science and Engineering, and Louis Block Fund of the University of Chicago (M.L.).

18 July 2001; accepted 9 November 2001

Role of Cell-Specific SpoIIIE Assembly in Polarity of DNA Transfer

Marc D. Sharp and Kit Pogliano*

SpoIIIE mediates postseptational chromosome partitioning in *Bacillus subtilis*, but the mechanism controlling the direction of DNA transfer remains obscure. Here, we demonstrated that SpoIIIE acts as a DNA exporter: When SpoIIIE was synthesized in the larger of the two cells necessary for sporulation, the mother cell, DNA was translocated into the smaller forespore; however, when it was synthesized in the forespore, DNA was translocated into the mother cell. Furthermore, the DNA-tracking domain of SpoIIIE inhibited SpoIIIE complex assembly in the forespore. Thus, during sporulation, chromosome partitioning is controlled by the preferential assembly of SpoIIIE in one daughter cell.

Table 1. Neutrality tests of haplotype structures in the *D. melanogaster* fourth chromosome genes.

Tests	Tested statistics	Observed values	Probability
K test	Haplotype number K	5	0.0435
H test	Haplotype diversity H	0.618	0.0055
HP test	Number of alleles with ≤ -1 segregating site	8	0.0050

The spore formation pathway of *Bacillus subtilis* provides a valuable system for studying how bacterial cells establish the cellular polarity necessary for development (1, 2). Early in sporulation, a polar septum is synthesized in the space between two domains of an asymmetrically partitioned chromosome (3). After divi-

sion, the forespore contains the origin proximal 30% of its chromosome, whereas the remaining 70% must subsequently be transported through the septum. This striking chromosome movement is accomplished by the SpoIIIE DNA translocase (4, 5), a bifunctional protein that also participates in membrane fusion after the phagocytosis-like process of engulfment (Fig. 1A) (6). The NH₂-terminal membrane domain of SpoIIIE is necessary and sufficient for localization to the septum, whereas the COOH-terminal domain moves along DNA in an adeno-

Division of Biology, University of California, San Diego, La Jolla, CA 92093–0349, USA.

*To whom correspondence should be addressed. E-mail: kpogliano@ucsd.edu