Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in *Drosophila melanogaster*

J. J. Emerson *et al.*

Science **320**, 1629 (2008);
DOI: 10.1126/science.1158078

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.

that improved this manuscript. Supported by NWO's Netherlands Polar Programme (M.M.H., W.J.v.d.B.) and by NASA's Cryospheric Sciences Program (C.H.D., Y.L.). ERS-2 radar altimeter data were provided by NASA Goddard Space Flight Center. All authors have discussed results and contributed to the manuscript. M.M.H. developed the firn densification model, and integrated the results. M.M.H., M.R.v.d.B., and R.S.W.v.d.W. frequently discussed results. M.R.v.d.B., W.J.v.d.B., and E.v.M. contributed to RACMO2/ANT data. C.H.D. and Y.L. analyzed elevation changes from ERS-2 data. I.G. contributed to accumulation records.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1153894/DC1
Materials and Methods
Figs. S1 to S9
Table S1
References

7 December 2007; accepted 7 May 2008
Published online 29 May 2008;
10.1126/science.1153894
Include this information when citing this paper.

Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in *Drosophila melanogaster*

J. J. Emerson,[1,2]*† Margarida Cardoso-Moreira,[1,3,4]*† Justin O. Borevitz,[1] Manyuan Long[1]

The role that natural selection plays in governing the locations and early evolution of copy-number mutations remains largely unexplored. We used high-density full-genome tiling arrays to create a fine-scale genomic map of copy-number polymorphisms (CNPs) in *Drosophila melanogaster*. We inferred a total of 2658 independent CNPs, 56% of which overlap genes. These include CNPs that are likely to be under positive selection, most notably high-frequency duplications encompassing toxin-response genes. The locations and frequencies of CNPs are strongly shaped by purifying selection, with deletions under stronger purifying selection than duplications. Among duplications, those overlapping exons or introns, as well as those falling on the X chromosome, seem to be subject to stronger purifying selection.

Differences in the numbers of copies of large DNA segments are an abundant source of genetic variation in humans (*1*, *2*), mice (*3*), and flies (*4*). Because CNPs can create new genes, change gene dosage, reshape gene structures, and/or modify the elements that regulate gene expression, understanding their evolution is at the very heart of understanding how such structural changes in the genome contribute to the phenotypic evolution of organisms (*5*–*7*).

A rigorous characterization of CNPs requires high-resolution data unbiased with respect to genome annotation. We used tiling arrays covering the full euchromatic genome of *D. melanogaster* at a median density of one unique perfect match probe for every 36 base pairs (bp) (*8*, *9*) in 15 natural isofemale lines (table S1). We inferred copy-number changes with a hidden Markov model (HMM) (*9*) that inferred the posterior probabilities for copy number by comparing DNA hybridization intensities between natural isolates and the reference genome strain. Training data for copy-number changes were obtained via hybridization with a line known to contain a ~200-kb homozygous duplication and from a set of 52 validated homozygous deletions (*9*). The probabilities of mutation were parsed to make CNP calls (table S3).

Because tiling arrays are restricted to nonredundant regions in the reference genome, deletion and duplication are detected by the absence of nonredundant DNA and by the doubling of unique DNA, respectively. In principle, it is possible to confound unique duplications with multiple hit scenarios of deletion of ancestral duplications. However, the few CNPs that exhibited even weak signs of ancestral redundancy in either *D. simulans* or *D. yakuba* (109 CNPs) showed a site-frequency spectrum (SFS) suggesting that the derived state cannot be a deletion [table S4; (*9*)]. Nevertheless, we excluded those events from our analyses.

In order to validate the CNP predictions, we performed polymerase chain reaction–based assays (*9*). For duplications, we obtained a false-positive rate of 14% and a false-negative rate of 16%. Notably, our assay can only amplify tandem duplications lying within several kilobases of each other, suggesting that the false-positive rate is overestimated. Conversely, the fact that we confirmed 86% of the duplications confirms that most CNPs form in tandem. For deletions, we obtained a false-positive rate of 47%. This high rate of falsely called deletions is in part due to the prevalence of multiple adjacent single-nucleotide polymorphisms (SNPs) in highly polymorphic regions of the *D. melanogaster* genome (*10*). We also obtained a false-negative rate of 18% for homozygous deletions and 32% for heterozygous deletions.

We detected 2658 unique CNPs among all 15 lines of *D. melanogaster*, with an average of 312 CNPs (SD = 31.9 CNPs), after adjusting for false positives. Except where noted, total mutation counts are corrected only for false positives. In total, CNPs comprise ~2% of the genome. The size distribution of CNPs was roughly exponential, with most being small variants (median: 336 bp) and few being larger variants (maximum size detected: 35 kb). The predicted and real CNP boundaries differ only by about one probe for duplications and about three probes for deletions (table S3). These data indicate that we were able to both detect

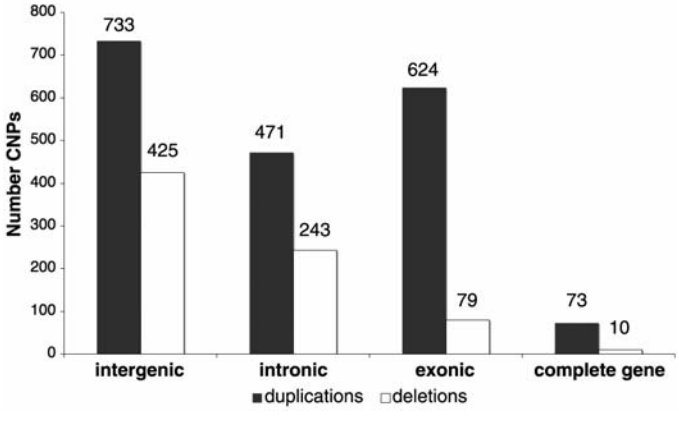

Fig. 1. Frequency of CNPs within different genomic contexts. The numbers of polymorphic duplications (black) and deletions (white) are shown for four mutually exclusive genomic contexts: intergenic (mutations between genes), intronic (mutations entirely within introns), exonic (mutations that overlap exons but not complete gene structures), and complete gene (mutations that overlap at least one complete gene structure, including UTRs).

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA. [2]Genomics Research Center, Academia Sinica, Taipei 115, Taiwan. [3]Graduate Program in Areas of Basic and Applied Biology, Universidade do Porto, Porto, Portugal. [4]Faculdade de Ciências, Universidade do Porto, Porto, Portugal.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jje@uchicago.edu (J.J.E.); mmoreira@uchicago.edu (M.C.M.)

Table 1. Description of the CNP dataset: number of events, frequency of singletons (CNPs detected in only one population), and size. We assumed a false-positive rate of 14% for duplications and 47% for deletions. Size and frequency of singletons were determined with the raw data.

CNP type	Number of events		Size (bp)		Frequency of singletons
	Raw data	Corrected false positives	Median	Mean	
Duplications	2211	1901	367	1117	0.75
Deletions	1428	757	282	604	0.67

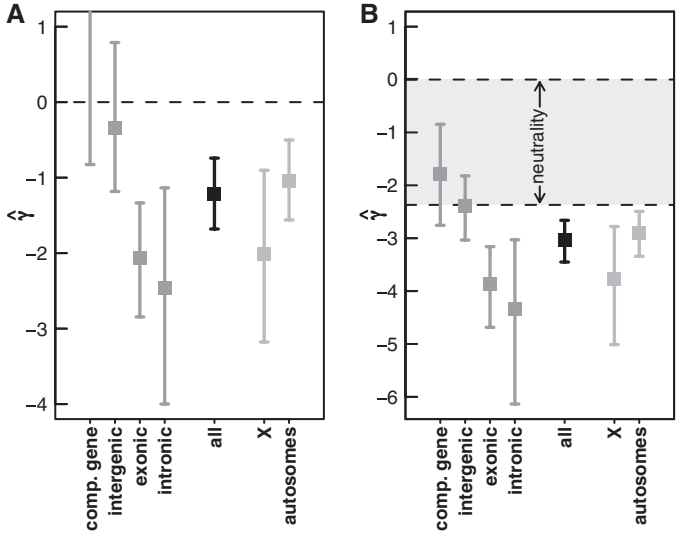

Fig. 2. Selection coefficients for polymorphic duplications with estimates obtained with the PRF-SFS methodology (*19*) are shown, both with and without incorporating ascertainment bias and error into the likelihood [(**A**) and (**B**), respectively]. Squares indicate the maximum likelihood estimates of γ (selection coefficient). Error bars indicate the 95% confidence interval ($\alpha > 0.05$ in a likelihood ratio testing framework) for the parameter γ. The upper bound was not plotted for complete (comp.) gene duplications because of low sample size (S = 67 for comp. genes). The gray region in (B), indicating neutrality, is bounded above by $\gamma = 0$ and below by γ_{sim}, which is estimated from simulations and corrected for ascertainment and error, for a neutral SFS expectation in the population of 10 strains [section 7.1.1 in (*9*)].

small CNPs as well as estimate CNP boundaries with precision (table S4). Despite a smaller sample size and a smaller genome, this study detected more CNPs than a recent survey in humans [2658 detected here versus 1447 detected in (*2*)]. This discrepancy is likely explained by the denser genome coverage in this study. Our data suggest that humans harbor a class of CNPs that is much larger than anything observed in fruit flies and that recent mammalian studies may be neglecting most small-scale variations.

Duplications outnumbered deletions 2.5:1 (Sign test P value $<2.22 \times 10^{-16}$; Fig. 1) and were significantly larger (Wilcoxon rank sum test, P value $<2.22 \times 10^{-16}$; Table 1). One mechanism thought to be an important contributor to tandem CNP formation—nonallelic homologous recombination—leads to either one gamete with a duplication and another with a complementary deletion or only one gamete carrying a deletion (*11*). Thus, nonallelic homologous recombination generates either an equal number of each mutation or an excess of deletions. Additionally, studies of insertion and deletion variation have shown a deletion bias in *D. melanogaster*, although the mutations' size (*12*) was considerably smaller than those examined here. The fact that we observed fewer deletions when either an equal number or an increased number of deletions was expected suggests that a large proportion of deletions are removed from the population by purifying selection. In this context, the dearth of deletions observed in our data, as well as the smaller size of the deleted variants, suggest that they are far more deleterious than duplications and that larger mutations are more deleterious than smaller ones.

Every region of the genome harbors at least low levels of CNPs. The median distance between two events was 12.6 kb (fig. S5). We found that pericentromeric regions were enriched in duplications, though not in deletions (fig. S5). Such regions are known to be rich in duplications (*13*). Redundancy results in a lower probe resolution in those regions, suggesting that our observation of increased levels of polymorphism was actually conservative. However, given the lower probe resolution in our work and the smaller size of deletions, we cannot assume that the absence of deletions in such regions is not artifactual. Pericentromeric regions are also characterized by extremely low rates of crossing-over, leading to a lower effective population size as a result of linkage (*14*). Therefore, the higher density of CNPs observed in these regions may be a consequence of the reduced effectiveness of selection in purging deleterious mutations (*14*). Alternatively, the mutation rate may simply be higher in such regions (*15*).

The genome distribution of CNPs varied significantly both between genome regions (i.e., coding versus noncoding) as well as between mutation types (i.e., duplication versus deletion) (Fig. 1). Duplications outnumbered deletions in all categories (all Sign test P values $<1 \times 10^{-10}$). Deletions falling in coding regions represented a smaller proportion of all deletions as compared with duplications (Fig. 1, Fisher's exact test P value $<2.2 \times 10^{-16}$).

Given the high incidence and widespread genomic distribution of CNPs, it is not surprising that 8 and 2% of genes were at least partially duplicated or deleted, respectively. Before correcting for false positives, we found 133 genes completely duplicated and 27 completely deleted (table S5). Among completely deleted genes, two have known, nonlethal mutant phenotypes (*16*). Tandem duplications of a sequence partially overlapping adjacent genes may create a chimera between them while leaving intact versions of both donor genes. We identified 92 CNPs that appear to be such chimeras. Curiously, 1.5 times as many duplications overlap the ends of genes than their starting points (Sign test P value = 0.0101), which is similar to the excess of transposable element insertions observed in 3′ untranslated regions (3′ UTRs) in *D. melanogaster* (*17*).

Taken together, the evidence above suggests that purifying selection eliminates a large fraction of standing CNP variation, especially deletions. Previous research on CNPs in humans (*1*) suggests that purifying selection may shape patterns of copy-number variation. Therefore, we tested selection on these variants in *D. melanogaster* by analyzing the distribution of allele frequencies (the SFS) [table S7 and fig. S8; (*18*)]. Purifying selection against deleterious mutations increases the fraction of rare variants, which is a common signature of natural selection. However, an excess of rare variants may also represent demographic processes such as population expansion, bottlenecks, or population structure (*19*). In order to quantify these effects, we sampled putatively neutral mutations. We collected ~600 synonymous SNPs from 46 loci located in all major chromosome arms in all 15 lines (*9*) and eliminated the effects of population structure (*9*, *20*). We then estimated demographic parameters for two models using a Poisson random fields–SFS (PRF-SFS) approach (*19*): (i) a two-epoch model to identify recent population expansions and (ii) a three-epoch model to identify bottlenecks (*21*–*23*). Because neither scenario rejected the neutral model ($P = 0.39$ and $P = 0.07$, respectively), we used the standard neutral model as the demographic null hypothesis (*9*). All SFS analyses were performed with raw CNP calls

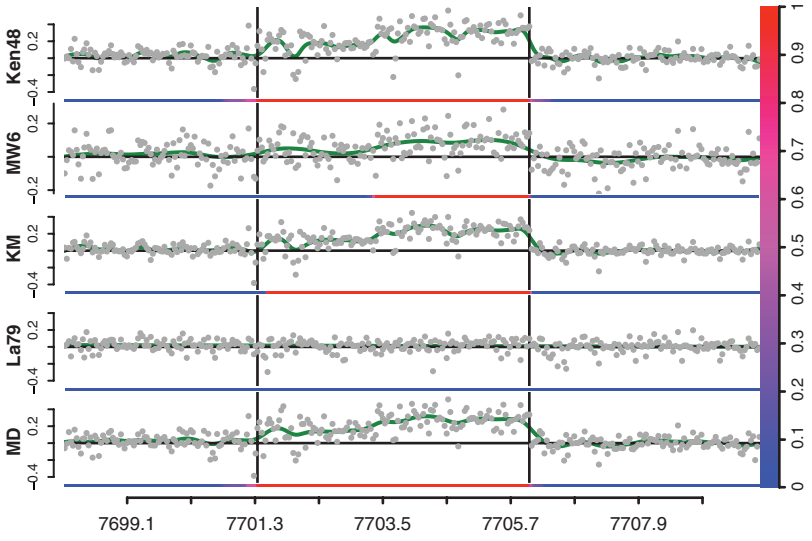

Fig. 3. Representation of a subset of 5 out of 15 individuals for the *Cpy6g1* polymorphism. The image in each row represents the log ratio of array intensities for the natural and reference lines as a function of genome position on chromosome 2R in kilobases. The green line is a smoothing spline for reference. The shading below each image indicates the posterior probability of duplication from the HMM, with red indicating a probability of 1 and blue indicating a probability of 0. The vertical lines indicate our boundary calls.

greater than 500 bp to restrict our inferences to mutations with smaller error rates, with error and bias corrected [as described in (*9*)].

We estimated γ, the scaled coefficient of natural selection (*9*, *19*). Our estimates show that natural selection is a pervasive force shaping the standing variation in *D. melanogaster* (Fig. 2). Notably, selection differentially influenced CNP evolution among different genomic features as well as among different chromosomes. We compared the patterns of variation between the different classes of variants: both correcting for bias and error and with no corrections. For inferences incorporating error and bias (Fig. 2A), we found that the intronic class exhibited the largest reduction in variation (γ = –2.5), although duplications within exons were only slightly less disfavored (γ = –2.1). We detected a significantly higher constraint in intronic than in intergenic regions (γ = –0.34). This observation contrasts with studies of nucleotide variation that found similar levels of constraint in both regions (*24*, *25*). This may be because introns are more strongly constrained by changes in size [e.g., for proper splicing (*26*, *27*)]. We hypothesize that duplications involving partial gene structures (the exonic and intronic classes) were the most strongly disfavored, because such mutations often result in the disruption of genes.

Notably, complete gene duplications showed the least constraint. Despite our conservative corrections for bias and error (*9*), we fail to reject neutrality. This unexpected observation is compatible with the hypothesis that full duplications are redundant. This result should, however, be interpreted with caution, because the synonymous SNPs that were used to parameterize the demographic model may be under weak purifying selection, potentially leading to an underestimate of the selection coefficient. Also, assuming a fixed selection coefficient may be wrong, because the set of complete gene duplications may include both advantageous and deleterious mutations.

We also found that the autosomes have higher selection coefficients than the X chromosome (Fig. 2). This observation is compatible with the following models: (i) duplicate mutations on the X chromosome are more deleterious than those on autosomes (X-linked genes may be more sensitive to changes in dosage) and/or (ii) duplicate polymorphisms tend to be slightly deleterious and recessive.

We identified five duplications overlapping seven genes involved in the response to toxins. For example, a duplication encompassing *Cyp6g1* and *Cyp6g2* was present in 13 of the 15 lines. *Cyp6g1* confers resistance to DDT and is known to be under positive selection for increased gene product [Fig. 3; (*28*)]. Three other independent high-frequency duplication events overlap four other genes (*Ugt86Dj*, *Ugt86Dh*, *CG30438*, and *CG10170*) involved in the response to toxins, and we found another duplicate gene (*Ugt86Di*, in one line) involved in the response to toxins. These duplications are good candidates to be under positive selection.

Overall, we present compelling evidence that the regional patterns of duplicate and deletion variation showed strong evidence for the pervasive action of natural selection, both in their patterns of polymorphism and in their distribution in the genome. These conclusions provide a comprehensive picture of the polymorphic phase of copy-number change.

References and Notes

1. L. Feuk, A. R. Carson, S. W. Scherer, *Nat. Rev. Genet.* **7**, 85 (2006).
2. R. Redon *et al.*, *Nature* **444**, 444 (2006).
3. T. A. Graubert *et al.*, *PLoS Genet.* **3**, e3 (2007).
4. E. B. Dopman, D. L. Hartl, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19920 (2007).
5. M. Long, E. Betrán, K. Thornton, W. Wang, *Nat. Rev. Genet.* **4**, 865 (2003).
6. S. P. Otto, P. Yong, *Adv. Genet.* **46**, 451 (2002).
7. F. A. Kondrashov, A. S. Kondrashov, *J. Theor. Biol.* **239**, 141 (2006).
8. F. Biemar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15907 (2005).
9. Materials and methods are available as supporting material on *Science* Online.
10. J. O. Borevitz *et al.*, *Genome Res.* **13**, 513 (2003).
11. D. J. Turner *et al.*, *Nat. Genet.* **40**, 90 (2008).
12. D. A. Petrov, E. R. Lozovskaya, D. L. Hartl, *Nature* **384**, 346 (1996).
13. A.-S. Fiston-Lavier, D. Anxolabéhère, H. Quesneville, *Genome Res.* **17**, 1458 (2007).
14. I. Gordo, B. Charlesworth, *Curr. Biol.* **11**, R684 (2001).
15. C. M. Bergman, H. Quesneville, D. Anxolabéhère, M. Ashburner, *Genome Biol.* **7**, R112 (2006).
16. M. Ashburner, R. Drysdale, *Development* **120**, 2077 (1994).
17. M. Lipatov, K. Lenkov, D. A. Petrov, C. M. Bergman, *BMC Biol.* **3**, 24 (2005).
18. B. Charlesworth, C. H. Langley, *Annu. Rev. Genet.* **23**, 251 (1989).
19. S. H. Williamson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7882 (2005).
20. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* **155**, 945 (2000).
21. P. R. Haddrill, K. R. Thornton, B. Charlesworth, P. Andolfatto, *Genome Res.* **15**, 790 (2005).
22. H. Li, W. Stephan, *PLoS Genet.* **2**, e166 (2006).
23. J. E. Pool, C. F. Aquadro, *Genetics* **174**, 915 (2006).
24. P. Andolfatto, *Nature* **437**, 1149 (2005).
25. D. L. Halligan, P. D. Keightley, *Genome Res.* **16**, 875 (2006).
26. S. M. Mount *et al.*, *Nucleic Acids Res.* **20**, 4255 (1992).
27. M. Deutsch, M. Long, *Nucleic Acids Res.* **27**, 3219 (1999).
28. P. J. Daborn *et al.*, *Science* **297**, 2253 (2002).

29. J.J.E. was supported by an NSF Graduate Research Fellowship, an NSF Doctoral Dissertation Improvement Grant, and a Graduate Assistantships in Areas of National Need training grant. M.C.M. was supported by the Portuguese Foundation for Science and Technology (POCI 2010, FSE). J.O.B. was supported by NIH R01GM073822. M.L. was supported by the Packard Fellowship for Science and Engineering, the NSF Career Award (MCB-0238168), and NIH (R0IGM065429-01A1 and R0IGM078070-01A1). We thank J. Pool and M.-L. Wu for providing the lines used in the study; J. Byrnes, G. Coop, R. Hudson, J. Shapiro, K. Thornton, R. Arguello, M. Vibranovski, and other members of the M.L. laboratory for discussions; A. Boyko for PRF-SFS methodology and code; M. Noe for characterizing the hybridization properties of the training line; and M.-Y. Lu, H.-M. Sung, and J. Spofford for help with the manuscript. Array information is deposited in the Gene Expression Omnibus database (accession number GSE11326), and sequence polymorphism information is deposited in the GenBank database (accession numbers EU706459 to EU707148).

Supporting Online Material

www.sciencemag.org/cgi/content/full/1158078/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 to S7
References and Notes

20 March 2008; accepted 15 May 2008
Published online 5 June 2008;
10.1126/science.1158078
Include this information when citing this paper.