Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*

Steffen Beisswanger* and Wolfgang Stephan

Section of Evolutionary Biology, BioCenter, University of Munich, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved February 6, 2008 (received for review November 16, 2007)

The *polyhomeotic* (*ph*) locus in *Drosophila melanogaster* consists of the two tandemly duplicated genes *ph-d* (distal) and *ph-p* (proximal). They code for transcriptional repressors belonging to the Polycomb group proteins, which regulate homeotic genes and hundreds of other loci. Although the duplication of *ph* occurred at least 25 million to 30 million years ago, both copies are very similar to each other at both the DNA and the protein levels, probably because of the action of frequent gene conversion. Despite this homogenizing force, differential regulation of both transcriptional units suggests that the functions of the duplicates have begun to diverge. Here, we provide evidence that this functional divergence is driven by positive selection. Based on resequencing of an ≈30-kb region around the *ph* locus in an African sample of *D. melanogaster* X chromosomes, we identified a selective sweep, estimated its age and the strength of selection, and mapped the target of selection to a narrow interval of the *ph-p* gene. This noncoding region contains a large intron with several regulatory elements that are absent in the *ph-d* duplicate. Our results suggest that neofunctionalization has been achieved in the *Drosophila ph* genes through the action of strong positive selection and the inactivation of gene conversion in part of the gene.

gene duplication | selective sweep

Gene duplication is a major evolutionary mechanism for generating new genes and new functions, thus increasing organismal complexity. Since Ohno's (1) pioneering ideas, this topic has become a main stay of evolutionary biology research. Ohta (2) laid the groundwork by exploring the population genetic mechanisms leading to the maintenance of gene duplications and the evolution of multigene families. The early evolution and functional divergence of duplicated genes, on the other hand, remained somewhat obscure. Theoretical work suggests that both genetic drift and positive selection may play a role in the fixation and early evolution of duplicate genes (3–5). In particular, positive selection is thought to drive the fixation of a duplicate gene that has gained a new function through acquisition of a beneficial mutation, a process referred to as neofunctionalization (6). However, there is currently limited evidence for this suggestion. The reasons for this may be twofold. First, it is difficult to detect newly diverging gene copies and, at the same time, identify selection at one and/or the other copy at the population level. Second, recent mathematical modeling predicts that neofunctionalization is unlikely in the presence of gene conversion, unless selection is very strong (7).

One possible method for detecting strong positive selection and localizing the target of selection is the search for selective sweeps. A selective sweep denotes a signature of variation in the genome that results from the recent fixation of a new, strongly selected beneficial mutation (8) or standing low-frequency variants (9). Such footprints last not much longer than $0.1\ N_e$ generations, where N_e is the effective population size (10, 11). In the past 5 years, several approaches have been proposed to detect sweeps in SNP data (10, 12, 13). The hallmark of a selective sweep is a severe reduction of genetic variation at the site of the beneficial mutation in the genome and an increase of nucleotide diversity at both sides of the selected site. The rate of increase depends on the ratio of the frequency of crossing-over to the strength of selection (8). Thus, if the rate of crossing-over is sufficiently high and the sweep was recent, this diversity-reducing feature of a sweep can be used to map the target of selection in the genome relatively accurately (potentially down to the gene level; refs. 14–16).

In a previous study, Beisswanger *et al.* (17) identified a genomic region in an African population (from the ancestral species range) that most likely has been affected by positive directional selection in the recent past. However, because of partial sequencing in this region, they were unable to identify the target of selection with high confidence. Interestingly, this region encompasses the locus *polyhomeotic* (*ph*) with the two tandemly duplicated copies *ph-p* (proximal) and *ph-d* (distal) that were in the approximate center of the swept segment. The *ph* genes code for transcriptional repressors belonging to the Polycomb group of proteins, which is known to regulate homeotic genes and also hundreds of other genes in mammals and insects (18). The distal protein is very similar to the proximal product, except for the absence of an amino acid terminal region and a small region near the carboxyl terminus (19). Over a stretch of 1.3 kb both proteins are completely identical. On the other hand, the first intron of *ph-p* is much larger than that of *ph-d*. With regard to functional divergence, there is no clear evidence that the distal and proximal products bind to and modify chromatin in different ways. However, other results (such as the differential regulation of both transcriptional units at the mRNA level) suggest that the functions of these proteins have begun to diverge (20). Thus, although the *ph* duplication is relatively old (see *Discussion*), the fact that the duplicates show little or no amino acid differences but significantly different expression patterns suggests that the two genes are still in an early phase of their sequence and functional divergence.

Here, we present results from a resequencing study of an ≈30-kb region around the *ph* genes, based on a sample of 12 *Drosophila melanogaster* X chromosomes. We report strong evidence for a selective sweep in this region and map the target of selection to *ph-p*, which indicates that positive selection drives the early sequence and functional divergence of the *ph* genes.

Results

Polymorphism in the *ph-d*–CG3835 Region. In a previous study we reported evidence for a selective sweep that affected the

Author contributions: S.B. and W.S. designed research; S.B. performed research; S.B. and W.S. analyzed data; and S.B. and W.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AM943662–AM943805).

*To whom correspondence should be addressed. E-mail: beisswanger@lmu.de.

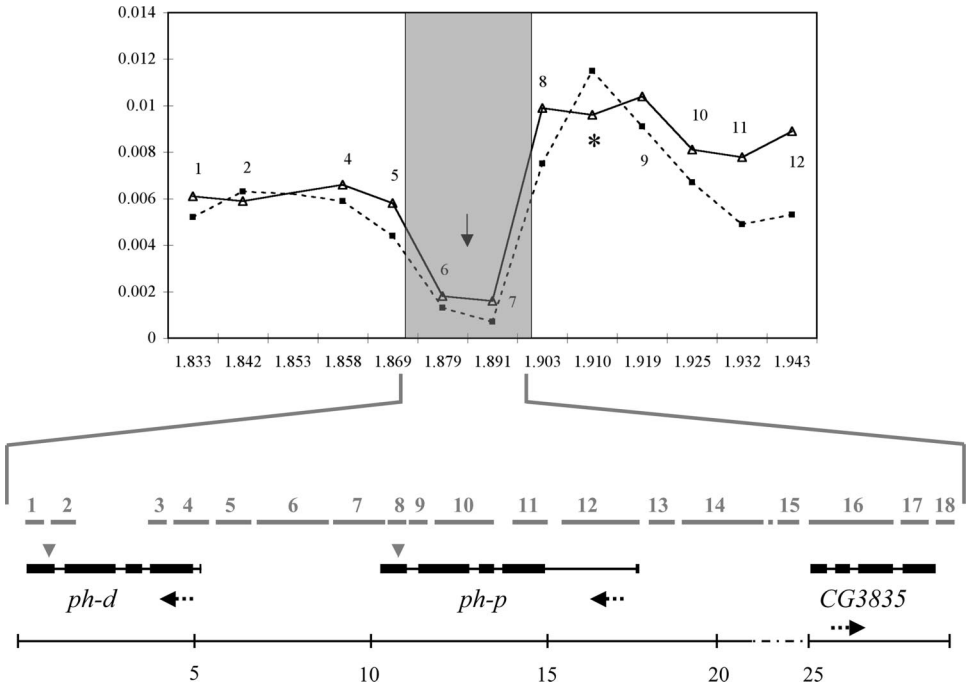

Fig. 1. Illustration of the *ph-d*–CG3835 region. (*Upper*) Modified from ref. 17, the entire *wapl* region is shown. The *wapl* fragment is indicated by a star. It harbors no variation in the European sample and was detected in a previous genome scan (34). The numbers on the *x* axis are the absolute genomic position (in Mb), and those on the *y* axis are nucleotide variability in the African sample (the solid line corresponds to θ and the dashed line corresponds to π). The arrow indicates the position of the target of selection predicted by Beisswanger *et al.* (17). (*Lower*) The region investigated in detail in this study. Exons are illustrated by thick black lines and introns by thin black lines. The direction of transcription is indicated by dotted arrows below each gene. Sequenced fragments are denoted by gray lines and numbers. Gray arrow heads indicate the position of the three SNPs shared between both *ph* copies. These SNPs are located within a region of 135 bp.

genomic region around the *wapl* gene of an African population of *D. melanogaster* (17). Several neutrality tests and the recently proposed composite likelihood ratio (CLR) test (10) pointed toward a target of selection located in the center of the region (within or near the gene *ph-p*). Therefore, we decided to sequence the *ph-d*–CG3835 region, i.e., the segment between the 3′ flanking region of *ph-d* and the 3′ flanking region of *CG3835* encompassing the gene *ph-p* (Fig. 1). This segment comprises 31,700 annotated base pairs (*D. melanogaster* genome release 4.3; ref. 21). However, small parts of the region could not be sequenced because of primer malfunction (see below). Fragment positions and individual summaries are provided in Table 1. In

Table 1. Summaries of SNP data

Locus	Position	*L*	*S*	*h*	*Hd*	Z_{nS}	θ	π	D_T	D_{FL}	*K*
1	0–515	515	3	4	0.68	—	0.0031	0.0026	−0.58	−1.12	0.05
2	623–1,109	486	2	3	0.32	—	0.0040	0.0020	−1.45	−0.48	—
3	3,562–4,051	489	1	2	0.17	—	0.0029	0.0015	−1.14	−1.42	0.03
4	4,111–5,198	1,087	7	6	0.76	1*	0.0027	0.0031	−0.87	−1.12	—
5	5,246–6,351	1,105	6	8*	0.89	0.11	0.0018	0.0015	−0.56	−1.07	0.06
6	6,492–8,493	2,001	15	11*	0.99*	0.24	0.0025	0.0017	−1.36	−1.78*	0.06
7	8,585–0,122	1,537	17	7	0.88	0.46	0.0037	0.0029	−0.98	−1.57	0.07
8	10,124–0,578	454	1	2	0.17	—	0.0018	0.0009	−1.14	−1.42	0.03
9	10,625–1,327	702	7	8	0.85	0.38	0.0076	0.0062	−1.07	−1.06	0.03
10	11,683–3,402	1,719	8	8	0.93	—	0.0044	0.0036	−1.06	−1.02	0.04
11	14,006–5,274	1,268	2	3	0.32	—	0.0008	0.0004	−1.45	−1.42	0.02
12	15,641–8,254	2,613	13	10*	0.96	—	0.0017	0.0009	−1.97**	−2.52**	0.04
13	18,317–8,967	650	5	5	0.58	—	0.0026	0.0013	−1.83**	−2.67**	0.04
14	19,061–23,280	4,219	41	12**	0.99**	0.24	0.0033	0.0021	−1.62*	−1.67*	0.09
15	23,338–23,958	620	16	9	0.91	0.27	0.0086	0.0087	0.05	−0.43	0.07
16	24,032–26,188	2,156	33	11*	0.99	0.12	0.0074	0.0047	−1.66*	−2.26*	0.05
17	26,258–26,814	557	2	2	0.3	1*	0.0052	0.0048	−0.25	0.95	0.03
18	27,155–27,372	220	2	2	0.3	1*	0.0030	0.0028	−0.25	0.95	0.05

Locus names and positions are according to Fig. 1. *L* is the length of a given locus; *S*, number of SNPs; *h*, number of haplotypes; *Hd*, haplotype diversity; Z_{nS}, LD; D_T, Tajima's *D*; D_{FL}, Fu and Li's *D*; *K*, divergence between *D. melanogaster* and *D. simulans*. For loci 2 and 4 we could not obtain outgroup sequences. θ and π were estimated for noncoding and synonymous sites.
*Significant at the 0.05 level.
**Significant at the 0.01 level.

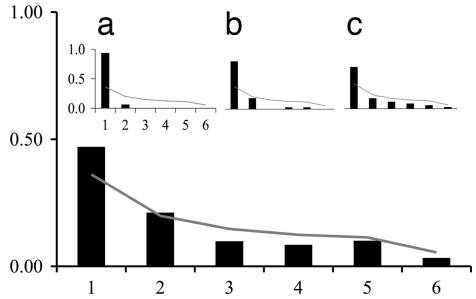

Fig. 2. Site frequency spectra. The folded frequency spectrum of 27 loci randomly chosen from ref. 23. (*Inset*) Shown are the spectra of the *ph-p* intron and the 5′ UTR (*a*), the first half of the *ph-p*–CG3835 intergenic region (*b*), and the second half of that region (*c*). The gray lines indicate the equilibrium frequencies.

total, we analyzed 18 fragments with an overall length of ≈27 kb (per line) in 12 African *D. melanogaster* lines.

Because *ph-d* and *ph-p* are paralogous genes (19, 22) sequencing was sometimes difficult. Hence, we blasted each sequenced fragment against the published *D. melanogaster* genome sequence and used only those fragments that resulted in a blast identity of >99.9%. We identified two segments where primers erroneously annealed to the paralogous gene. These fragments were excluded from subsequent analyses. Excluding sequence overlaps and indels, a total of 27,372 bp were used for further investigation.

As shown in Table 1, levels of nucleotide variability across the newly sequenced region are heterogeneous, but low, particularly in the vicinity of *ph-p*. For noncoding and synonymous sites average θ across the region is 0.0037. This value is lower than the chromosome-wide average ($\theta = 0.0126$; 23). In addition, average θ is ≈40% lower than the average level of nucleotide diversity reported for the 12 loci around the *ph-d*–CG3835 region (see ref. 17 and Fig. 1). In total, we observed 181 SNPs in our dataset. Here, we focus on noncoding sequences.

Haplotype Structure. The number of haplotypes across the region ranges from 2 to 12 with an average of 6 haplotypes per locus. Haplotype diversity (*Hd*) ranges from 0.17 to 1 (Table 1). Linkage disequilibrium (LD), as measured by Kelly's Z_{nS} statistic (24), is low to intermediate (average: 0.48), and only three values were significant (Table 1). Two of these loci (fragments 4 and 17) represent coding sequences. For eight loci no Z_{nS} estimates could be obtained, as all polymorphic sites were noninformative. Five of these fragments, i.e., loci 8 and 10–13, are located in *ph-p* and its 5′ flanking region. We also estimated LD in terms of r^2 (25) for all pairwise SNPs across the whole region. Only 14 significant associations were observed, of which 8 are located within *ph-d*, 5 in the *ph-d*–ph-p intergenic region, and 1 in the *ph-p*–CG3835 intergenic region. Assuming the selected site was estimated correctly by Beisswanger *et al.* (17), this result is in qualitative agreement with Stephan *et al.* (26), who predicted that after a hitchhiking event LD is reduced between polymorphisms across the selected site (relative to LD between pairs of polymorphisms located on the same side of the target of selection). Because *Hd* estimates were generally high (on average 0.67), we were also able to determine the minimum number of recombination events. A minimum of 23 crossovers was identified across the region investigated.

Standard Neutrality Tests and Frequency Spectra. As shown in Table 1, all fragments under investigation show a trend toward a negative Tajima's *D* (27), except fragment 15 for which $D = 0.05$. We obtained significantly negative estimates for four loci in the region analyzed ($D = -1.97, -1.83, -1.62$, and -1.66 for fragments 12, 13, 14, and 16, respectively). Three of these fragments are located in or immediately around *ph-p* (Fig. 1). Similarly, Fu and Li's *D* (28) was significantly different from zero ($P < 0.05$) for these loci ($D = -2.52, -2.67, -1.67$, and -2.26). This finding indicates an excess of low-frequency variants. We also applied a multilocus HKA test (29) to our data and obtained a significant result ($X^2 = 17.87$, $P = 0.037$). In addition, we analyzed our data in combination with the data of Beisswanger *et al.* (17). The result obtained by this approach was also significant ($X^2 = 32.79$; $P = 0.012$), which indicates that the *ph-d*–CG3835 region is not evolving according to the neutral equilibrium model. We note that in the presence of gene conversion (which may have played a role in the evolution of the *ph* genes; see below), the tests we have used here are conservative (30).

With regard to the site frequency spectrum, we observed a strong deviation from standard neutral expectations at *ph-p*. In a >4-kb stretch (encompassing the large intron of *ph-p*, its 5′ UTR and 417 bp of the *ph-p*–CG3835 intergenic region), we detected only singletons, with the exception of a doubleton at position 15,785 (Fig. 2 *Inset*, *a*). A similar, although less extreme, pattern is evident in the *ph-p*–CG3835 intergenic region. Here, the frequency spectrum of the first 40 segregating sites (corresponding to the first half of that region) also reveals an excess of low-frequency variants (Fig. 2 *Inset*, *b*). Then, going further away from *ph-p* toward *CG3835*, the frequency spectrum of the second half of 40 SNPs (Fig. 2 *Inset*, *c*) approaches that of the entire X chromosome (see Fig. 2).

Application of the CLR Test and the Strength of Selection. We applied the CLR test of Kim and Stephan (10) to our data to examine two different scenarios. First, we tested our new data (i.e., the *ph-d*–CG3835 region), assuming an expected nucleotide diversity of $\theta = 0.0037$, which is the average across the region. Second, we tested our dataset assuming $\theta = 0.0067$, which is the mean across the entire *wapl* region (17). As shown in Table 2, we obtained highly significant results for both scenarios ($P < 0.01$). Following Jensen *et al.* (31), our results were also robust against simple demographic scenarios ($P > 0.9$). In other words, positive selection rather than demography is the likely cause of the pattern observed. Estimates of the strength of selection (α) are 1,141 and 2,608 for both scenarios tested, which suggests that selection at *ph* was strong. Because the method for the estimation of α assumes that the sweep ended at present time (10), which

Table 2. Likelihood analysis of selective sweep

Parameter	$\theta = 0.0037$	*P*	95% C.I.	$\theta = 0.0067$	*P*	95% C.I.
LR	15.26	0.003		76.62	< 0.001	
$\hat{\alpha}$	1,141			2,608		
$\hat{X}$	15,224		9,267–22,110	15,159		8,130–21,449
Λ_{GOF}	1,272	> 0.9		1,273	> 0.9	

LR, likelihood ratio (see ref. 10); $\hat{\alpha}$, estimate of the strength of selection; $\hat{X}$, estimated position of the target of selection; Λ_{GOF}, estimate of the goodness-of-fit test statistic (31).

is not the case here (see below), the strength of selection is somewhat underestimated.

Target of Selection. The CLR test (10) also provides an estimate of the target of selection. For both scenarios the likelihood ratio was maximized at 15.2 kb. This position corresponds to the first intron of the gene *ph-p*, which is 4,291 bp in length according to release 4.3 of the *D. melanogaster* genome (21). Because the 1,445-bp transposable element (TE) FB{}23 is annotated in this region (positions 15,516 to 16,961 in our dataset), we tested all *D. melanogaster* lines and *Drosophila simulans* for the presence of the TE by diagnostic PCR. We could not detect any insertion in this genomic region in the *Drosophila* lines investigated (data not shown), which indicates that the first *ph-p* intron is actually 1,445 bp smaller than annotated.

We scrutinized this intron and the adjacent 5′ flanking region (noncoding) for fixed differences between *D. melanogaster* and its sibling species *D. simulans*, and *Drosophila yakuba*. In addition, we analyzed divergence between species. In parts of the intron and at the beginning of the 5′ flanking region we detected relatively little divergence among species. However, two segments of the intron, located at 16.3–16.6 kb and 17.3–17.7 kb, show high levels of divergence ($\geq 10\%$ to *D. simulans* and $>25\%$ to *D. yakuba*). A similar pattern can be observed in the 5′ flanking region. In the exons we could not detect any significant changes, except for a few single amino acid substitutions. However, we found indel differences between species in both the intron and 5′ region. There are two deletions of 28 and 37 bp in comparison with *D. simulans* (*D. yakuba*) at intron positions 15,747 and 17,226, respectively. In addition, a fixed 6-bp deletion can be found at position 18,925.

We also searched the first *ph-p* intron and the 5′ flanking region in the three Drosophilids for putative transcription factor binding sites (TFBs). We detected several putative TFBs with high confidence ($P > 0.95$), which are located in the first intron and the 5′ flanking region. Several of these are shared between *D. simulans* and *D. yakuba* but not between these species and *D. melanogaster*. These TFBs are located at the 5′ end of the intron or the proximate part of the 5′ flanking region. Interestingly, all of the detected TFBs are associated with the regulation of homeotic genes. For instance, at position 18,801 a GAGA factor binding site is predicted for *D. melanogaster*, but not for *D. simulans* and *D. yakuba*. At position 18,519 a *fushi tarazu* binding site is predicted for *D. melanogaster*, whereas a *caudal* binding site is found for the other two species. This difference is caused by an A to T substitution in the *D. melanogaster* lineage. Similarly, at intron position 17,383 a *retained* TFB is predicted for *D. melanogaster*, but not for *D. yakuba* and *D. simulans*, because of a nucleotide substitution.

Age of Selective Sweep. We estimated the age of the selective sweep by using two different approaches. First, we applied the method of Przeworski (11) to ≈4 kb of the *ph-p* 5′ flanking region. We assumed a distance of either 1 bp, 1.7 kb, or 3.2 kb to the selected site. These values correspond to a putatively selected site located in either the proximate 5′ flanking region (at the proximate end of fragment 14), the 5′ end of the *ph-p* intron, or the middle of that intron. For all distances tested this method returned age estimates between 45,000 and 55,000 years (assuming 10 generations per year; Table 3). Very similar values are obtained if ≈ 4 kb of the downstream region of the *ph-p* intron are used. We note that these estimates are surprisingly consistent. For both regions analyzed, the lowest values are obtained if the selected site is assumed to be in the 5′ flanking region.

Second, we estimated the time since the fixation of the beneficial allele following Slatkin and Hudson (32) and Ayala *et al.* (33). This method assumes a star-like genealogy, i.e., a complete selective sweep and subsequent accumulation of new mutations. We applied the method to two regions. First, we considered the first *ph-p* intron, which harbors 13 segregating sites in 2,613 bp, of which 12 are singletons. Assuming a local mutation rate of 0.9×10^{-9} per year (estimated from $K = 0.041$), this method yields an age estimate of 46,086 years. In addition, we applied this method to polymorphism data from 4,219 consecutive sites of the 5′ flanking region where 41 SNPs were observed. We obtained an estimate of 44,990 years since fixation assuming $\mu = 1.8 \times 10^{-9}$ ($K = 0.085$). Similar values were found for various stretches of the center of the sweep region.

In the presence of gene conversion, the first method should overestimate the age of the sweep because this process is expected to shift the frequency spectrum of polymorphic sites to intermediate values (30), whereas the second method (which is based on the number of segregating sites) should be less sensitive. The fact that both approaches returned comparable values (with regard to the method and the region analyzed) suggests that gene conversion does not introduce a major bias on our age estimates.

Table 3. Age of selective sweep (according to ref. 11)

Locus	Position	Distance	*T*	95% C.I.
14	19,061–23,280	1 bp	44,926	24,570–207,543
		1.7 kb	54,954	24,971–230,347
		3.2 kb	55,016	25,105–216,366
8–11	10,124–15,274	500 bp	54,708	20,309–524,159
		2 kb	55,301	20,025–456,477
		3.8 kb	45,155	20,433–420,838

Locus names and positions are according to Fig. 1. Distance is the interval from the edge of the locus analyzed to the presumed target of selection. *T*, estimate of the time since the fixation of the beneficial allele (in years).

Discussion

Evidence for a Selective Sweep in the *ph-p* Gene Region of African *D. melanogaster*. By resequencing ≈80% of an ≈30-kb region around the *ph* genes in a sample of 12 African *D. melanogaster* X chromosomes, we found strong evidence for a selective sweep in this region. Furthermore, our mapping showed that the target of selection is most likely located at the 5′ end of the large intron of *ph-p* or the proximate 5′ flanking region (within a region of ≈3 kb). We arrived at this conclusion after two rounds of mapping. First, we selected from an initial genome scan (34) a region with low variation (the *wapl* fragment) and found evidence for a sweep by resequencing several short fragments around *wapl* in a European and an African sample (17). For the European population a very large valley of very low variation (of ≈60 kb) was detected, whereas in the African population the valley was only ≈15 kb wide. Statistical tests showed evidence for a sweep in both populations. However, it was also plausible that the European pattern of variation was caused by a sweep that originated in Africa. For these reasons we decided to use the African population for a second round of mapping. Our estimates of the age of the sweep now support this rationale: the sweep occurred ≈50,000 years ago, most likely before the European *D. melanogaster* lineage split off from the African one (≈16,000 years ago; ref. 35). The estimated age of the sweep is also consistent with the observation that the sequences of the *ph-p* alleles from our European sample are entirely identical with that of the African haplotype in the swept region (data not shown).

It is clear that the evidence we provided is subject to uncertainty. First, the test of Kim and Stephan (10) for a selective sweep does not take the proper demographic model into account. However, we corrected for this by using the approach of

Jensen *et al.* (31). Furthermore, we applied the test in a conservative fashion using a low (local) estimate of θ (instead of the chromosome-wide value reported in ref. 23). The 95% C.I. around the selected site are much smaller than in the first mapping study, but still considerably large (although 80% of the *ph* region was sequenced). However, the frequency spectra (Fig. 2) support our conclusion that the most likely target of selection is located in a region of only a few kilobases between the 5′ end of the large *ph-p* intron and the proximal half of the *ph-p*–CG3835 intergenic region. Finally, although the 95% C.I. of the age of the sweep are also relatively large, the lower bound corresponds to an age of >20,000 years, suggesting that the sweep is older than *D. melanogaster*'s migration out of Africa. A more serious problem may be that the estimation method is based on an equilibrium population (11), which is probably not the case for the African population (35). Thus, taking into account the evidence that the African population increased (13) may change our estimates. However, two observations argue against this: first, according to the estimation of Li and Stephan (13) the last major population size expansion in Africa occurred ≈60,000 years ago, and second the method by Slatkin and Hudson (32), which is independent of N_e, yields results that are consistent with the estimates produced by Przeworski's (11) method.

Can polymorphism patterns at *ph-p* be explained by other factors? Functional elements within genes tend to be located within the first intron (36). In addition, it has been observed that long introns are more likely to harbor functional elements than short ones (37). It has therefore been argued that intron sequence evolution is more constrained than synonymous site evolution (38). This is partially reflected in the first *ph-p* intron, which shows only low to intermediate divergence between *D. melanogaster* and *D. simulans*. However, the amount of polymorphism at *ph-p* is severely reduced over >3 kb and results of common summary statistics (Table 1) observed at *ph-p* are more extreme than generally observed for conserved noncoding sequences (39). This suggests that purifying selection is not a likely cause of the observed reduction of variation.

Significance of the Selective Sweep and Neofunctionalization in the *ph* Genes. Which nucleotide site was the target of selection? Despite the relatively accurate mapping of the selective target, this question is difficult to answer. It may even be possible that several molecular variants within the small target region (of ≈3 kb) we identified were under positive selection, as this region contains several putative TFBs that are new in *D. melanogaster* or show differences between *D. melanogaster* and both *D. simulans* and *D. yakuba*.

Although the exact selective target(s) cannot be identified, the fact that the target region of selection contains new (or altered) TFBs relative to *D. simulans* and *D. yakuba* suggests that it is involved in the functional divergence of the two *ph* genes, which are differentially regulated (20). Furthermore, the evidence of a selective sweep that has been mapped to this region indicates that the functional divergence of these genes is driven by strong positive selection.

The observation that the distal and proximal Ph proteins are very similar and bind to and modify chromatin in similar ways, but that the transcriptional units of *ph-d* and *ph-p* are differentially regulated suggests that divergence of these two genes is still in its early stage. This is surprising, as the *ph* duplication likely occurred before the split of the *obscura* and *melanogaster* groups: a comparison of the 12 sequenced *Drosophila* genomes (40) shows that duplicated *ph* genes are already present in all species of the *melanogaster* group and *Drosophila pseudoobscura*. Thus, the duplication appears to be at least 25 million to 30 million years old (41). Given this relatively old age, it is interesting that the *ph* genes show little sequence divergence (5% for coding regions that could be aligned), and in fact, are completely identical at the protein level over a distance of ≈1.3 kb (19). In *D. pseudoobscura* sequence divergence between the *ph* duplicates is even smaller (< 3% for coding regions). There are two possible explanations for this. First, Ph is a member of the Polycomb group of proteins, which form very conserved protein complexes, suggesting that purifying selection at the *ph* locus is strong. This finding is in agreement with the generally low divergence at these genes (Table 1). Second, the identity of both genes may be maintained by frequent gene conversion. The number of segregating sites is too small to test the latter prediction (for instance, using the method described in ref. 42). However, there is some evidence for gene conversion at *ph* as a few polymorphic sites are shared between the duplicates (3 of 28 SNPs), which may suggest that the forces counteracting the functional divergence of the *ph* genes are relatively severe and can only be overcome by very strong positive selection (7). However, why was neofunctionalization eventually achieved? The likely reason appears to be that gene conversion was at some point inactivated in the first introns of *ph-p* and *ph-d*, because of drastic sequence differences caused by insertion/deletion events (43). As a result, the first intron of *ph-p* (which is much larger than that of *ph-d*) contains several regulatory elements that may be potential targets of selection (in that they can escape the counteracting force of gene conversion). We have shown here that some of these potential target sequences are located in precisely the narrow region of the genome that experienced recent strong positive selection. Taken together, these results provide consistent evidence for the process of neofunctionalization in duplicated genes under concerted evolution.

Materials and Methods

Fly Samples and Data. We collected data from 12 highly inbred *D. melanogaster* lines derived from an African population (44). For interspecific comparisons we used the publicly available sequences of *D. simulans*, *D. yakuba*, and *D. pseudoobcura* (40). For generating the frequency spectrum (Fig. 2), we randomly chose 27 putatively neutrally evolving fragments from ref. 23: loci 57, 60, 78, 84, 93, 106, 166, 178, 205, 206, 231, 237, 241, 259, 272, 276, 277, 286, 326, 359, 392, 422, 431, 439, 465, 721, and 727.

Molecular Techniques. Sequences of the region between the annotated genes *ph-d* and *CG3835* (the *ph-d*–CG3835 region) were generated as described (17), except that sequences were run on an ABI 3730 DNA Analyzer (Applied Biosystems). For alignments we used SeqMan and MegAlign (DNAstar). The ≈30-kb region under investigation was sequenced in overlapping fragments.

Statistical Analysis. Basic analyses were done by using DnaSP 4.10 (45). Nucleotide diversity was estimated in terms of θ (46) and π (47) using the number of haplotypes and Hd (48). Indels and singletons were not included in the analysis. We estimated the minimum number of recombination events in the whole region applying the four-gamete rule as implemented in LDhat (49). We tested the neutral equilibrium model by using Tajima's *D* (27), Fu and Li's *D* (28), and the HKA test (29). Significance was inferred by 10,000 neutral coalescent simulations under the assumption of equilibrium and zero recombination. Tests were performed as one-sided.

Likelihood Analysis of Selective Sweep. We analyzed our SNP data by using the CLR test of Kim and Stephan (10). We tested the significance of the resulting likelihood ratio by 10,000 Monte Carlo simulations under neutrality. The CLR test result was evaluated by the goodness-of-fit test (31). This test was also used to obtain C.I. for $\hat{X}$ by parametric bootstrapping. For the likelihood analyses we assumed $\theta = 0.0037$ and 0.0067 (instead of the chromosomewide estimate of $\theta = 0.0126$; ref. 23), which correspond to the average level of nucleotide diversity observed in the *ph-d*–CG3835 region and the entire *wapl* region (17), respectively. These lower θ values make the test more conservative.

Identifying the Target of Selection. After the identification of a putative selective sweep and localizing the approximate position of the selected site in the African sample, we screened the proximity of that site for candidate fixations that occurred along the *D. melanogaster* lineage. *D. simulans* and *D. yakuba* were used as outgroups. We also tested whether candidate fixations

EVOLUTION

in intronic and 5′ flanking regions result in regulatory changes (i.e., differences in TFBs) by applying the MatInspector tool (50).

Age of Selective Sweep. We used two methods to estimate the age of the selective sweep. First, we ran the algorithm of Przeworski (11) to obtain 2,000 successful matches. The time since the fixation of the selected site was determined by finding the mode of a histogram. Parametrization of the prior distributions is as follows: $N_e = 2.9 \times 10^6$, the mean mutation rate $\mu = 1.45 \times 10^{-9}$ per site per generation (assuming 10 generations per year), and crossing-over rate $\rho = 0.48 \times 10^{-8}$. The first two estimates are inferred from the entire X chromosome dataset (13), whereas the last one is for the (local) *wapl* region under consideration (17).

Second, we assumed a star-like genealogy subsequent to the sweep (for the region surrounding the presumed selected site) and computed the time since the sweep as T = number of SNPs / (number of sites × sample size × mutation rate per year) (32, 33). Local mutation rates μ were estimated as $\mu = K/(2t)$, where K is the local divergence between *D. melanogaster* and *D. simulans*, and t is the divergence time between the two species. We assumed $t = 2.3$ million years (13, 51).

ACKNOWLEDGMENTS. We thank Anne Wilken and Simone Lange for excellent technical assistance; David De Lorenzo, Hideki Innan, Pleuni Pennings, and Molly Przeworski for valuable suggestions; Tim Sackton for help with the *Drosophila* assemblies; and Neel Randsholt for introducing us to *polyhomeotic*. This work was supported by the Deutsche Forschungsgemeinschaft Grant STE 325/7.

1. Ohno S (1970) *Evolution by Gene Duplication* (Springer, Berlin).
2. Ohta T (1980) *Evolution and Variation of Multigene Families: Lecture Notes in Biomathematics 37* (Springer, Berlin).
3. Clark AG (1994) Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci USA* 91:2950–2954.
4. Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.
5. Walsh JB (2003) Population-genetic models of the fates of duplicate genes. *Genetica* 118:279–294.
6. Walsh JB (1995) How often do duplicated genes evolve new functions. *Genetics* 139:421–428.
7. Innan H (2003) A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc Natl Acad Sci USA* 100:8793–8798.
8. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genet Res* 23:23–35.
9. Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci USA* 101:10667–10672.
10. Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765–777.
11. Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. *Genetics* 164:1667–1676.
12. Nielsen R, *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res* 15:1566–1575.
13. Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* 13:e166.
14. Nurminsky DI, De Aguiar D, Bustamante CD, Hartl DL (2001) Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. *Science* 291:128–130.
15. Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc Natl Acad Sci USA* 101:1626–1631.
16. Orengo DJ, Aguadé M (2007) Genome scans of variation and adaptive change: Extended analysis of a candidate locus close to the *phantom* gene region in *Drosophila melanogaster*. *Mol Biol Evol* 24:1122–1129.
17. Beisswanger S, Stephan W, De Lorenzo D (2006) Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics* 172:265–274.
18. Schwartz YB, Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programs. *Nat Rev Gen* 8:9–22.
19. Deatrick J, Daly M, Randsholt NB, Brock HW (1991) The complex genetic locus polyhomeotic in *Drosophila melanogaster* potentially encodes two homologous zinc-finger proteins. *Gene* 105:185–195.
20. Hodgson JW, *et al.* (1997) The polyhomeotic locus of *Drosophila melanogaster* is transcriptionally and posttranscriptionally regulated during embryogenesis. *Mech Dev* 66:69–81.
21. Grumbling G, Strelets V, The FlyBase Consortium (2006) FlyBase: Anatomical data, images and queries. *Nucleic Acids Res* 34:D484–D488.
22. Dura J-M, *et al.* (1987) A complex genetic locus, polyhomeotic, is required for segmental specification and epidermal development in *D. melanogaster*. *Cell* 51:829–839.
23. Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the impact of demography and selection on *Drosophila melanogaster* from a chromosome-wide DNA polymorphism study. *Mol Biol Evol* 22:2119–2130.
24. Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics* 146:1197–1206.
25. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theor Appl Genet* 38:226–231.
26. Stephan W, Song YS, Langley CH (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* 172:2647–2663.
27. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
28. Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
29. Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
30. Innan H (2003) The coalescent and infinite-site model of a small multigene family. *Genetics* 163:803–810.
31. Jensen JD, *et al.* (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170:1401–1410.
32. Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.
33. Ayala FJ, Balakirev ES, Saez AG (2002) Genetic polymorphism at two linked loci, *Sod* and *Est-6*, in *Drosophila melanogaster*. *Gene* 300:19–29.
34. Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multilocus approach. *Genetics* 165:1269–1278.
35. Stephan W, Li H (2007) The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98:65–68.
36. Majewski J, Ott J (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res* 12:1827–1836.
37. Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P (2005) Patterns of intron evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol* 6:R67.
38. Halligan DL, Keightley PD (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res* 16:875–884.
39. Casillas S, Barbadilla A, Bergman CM (2007) Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol Biol Evol* 24:2222–2234.
40. *Drosophila* 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila phylogeny*. *Nature* 450:203–218.
41. Russo CAM, Takezsaki N, Nei M (1995) Molecular phylogeny and divergence times of Drosophilid species. *Mol Biol Evol* 12:391–404.
42. Innan H (2002) A method for estimating the mutation, gene conversion, and recombination parameters in small multigene families. *Genetics* 161:865–872.
43. Teshima KM, Innan H (2008) Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*, 10.1534/genetics.107.082933.
44. Begun DJ, Aquadro CF (1993) African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* 365:548–550.
45. Rozas J, Sánchez-Del Barrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
46. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* 7:256–276.
47. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.
48. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York), pp 177–178.
49. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination rates from gene sequences. *Genetics* 160:1231–1241.
50. Cartharius K, *et al.* (2005) MatInspector and beyond: Promoter analysis based on transcription factor binding sites. *Bioinformatics* 21:2933–2942.
51. Li Y-J, Satta Y, Takahata N (1999) Paleo-demography of the *Drosophila melanogaster* subgroup: Application of the maximum-likelihood method. *Genes Genet Syst* 74:117–127.