Recent Strong Positive Selection on *Drosophila melanogaster HDAC6*, a Gene Encoding a Stress Surveillance Factor, as Revealed by Population Genomic Analysis

Nicolas Svetec, Pavlos Pavlidis, and Wolfgang Stephan

Department of Biology, Section of Evolutionary Biology, LMU BioCenter, Planegg-Martinsried, Germany

Based on nearly complete genome sequences from a variety of organisms, data on naturally occurring genetic variation on the scale of hundreds of loci to entire genomes have been collected in recent years. In parallel, new statistical approaches (such as the composite likelihood ratio and "SweepFinder" tests) have been developed to infer evidence of recent positive selection from these data and to localize the target of selection. Here, we apply these methods to the X chromosome of *Drosophila melanogaster* in an effort to map genes involved in ecological adaptation. Using slight modifications of these tests that increase their robustness against past demographic changes, we detected evidence that recent strongly positive selection has been acting on a 2.7-kb region in an ancestral African population. This region overlaps with the 3′ end of *HDAC6*, a gene that encodes a newly characterized stress surveillance factor. HDAC6 is an unusual histone deacetylase being localized in the cytoplasm. Its ubiquitin-binding and tubulin-deacetylase activities suggest that HDAC6 is very different from other histone deacetylases. Indeed, recent discoveries have shown that HDAC6 is a key regulator of cytotoxic stress resistance.

Introduction

Recent advances in population genomics allow us to detect footprints of strong positive selection in the genome and to identify the targets of selection on the scale of individual genes (reviewed in Pavlidis et al. 2008). Based on nearly complete genome sequences from a variety of organisms, data on naturally occurring DNA sequence variation for hundreds of loci to entire genomes have been collected in the past 5 years. Most of these studies concentrated on *Drosophila melanogaster* (e.g., Glinka et al. 2003; Orengo and Aguadé 2004) and humans (e.g., Akey et al. 2004; International HapMap Consortium 2007). In parallel, new statistical tests have been developed to infer evidence of recent positive selection from these data (Kim and Stephan 2002; Jensen et al. 2005; Nielsen et al. 2005).

These new tests are based on the hitchhiking model developed by Maynard Smith and Haigh (1974). When a beneficial mutation arises in a population and goes to fixation driven by positive selection ("selective sweep"), theory predicts the emergence of a specific polymorphism pattern: 1) diversity vanishes around the site of selection, 2) the site frequency spectrum (SFS) of polymorphisms is shifted toward low and high frequency–derived variants (Braverman et al. 1995; Fay and Wu 2000), and 3) linkage disequilibrium (LD) is elevated in the early phase of the fixation process (Kim and Nielsen 2004; Stephan et al. 2006). Importantly, the width of the valley of reduced variation is mainly determined by the ratio of the rate of recombination around the site of selection and the strength of selection.

A multitude of studies has used the predictions of the hitchhiking model to detect footprints of positive selection in the genome of various organisms, estimate the strength of selection, and map the target of selection (Pavlidis et al. 2008). However, demographic factors such as population size bottlenecks may stochastically produce patterns of nucleotide diversity across the genome that resembles those of selective sweeps. Therefore, a major challenge of these analyses has been (and still is) to distinguish the effects of selection from those of demography. Recent progress in this area of research could be made based on the insight that demography affects the entire genome, whereas selection acts on individual loci. This improved the robustness of the tests for selection (Jensen et al. 2005; Nielsen et al. 2005; Thornton and Jensen 2007).

The candidate regions of selection identified by these tests, however, were generally very large (often ∼100 kb) and contained many genes (∼10). This is particularly the case for humans (e.g., Williamson et al. 2007). Although some progress has been made in *Drosophila* (Pool et al. 2006; Jensen et al. 2007; Orengo and Aguadé 2007; Beisswanger and Stephan 2008), a major challenge ahead is to develop strategies that help to narrow down the target regions of selection such that it is possible to map the site of selection to individual genes or gene regions. This is essential for ultimately understanding adaptation at the functional level.

Here, we use selection mapping to identify genes in *D. melanogaster* that may have been involved in ecological adaptation. We were able to identify a 2.7-kb region as the putative target of selection that contains the last exon of *HDAC6* harboring a ubiquitin-binding domain. HDAC6 is an unusual histone deacetylase with two catalytic domains and is localized in the cytoplasm. Its activities (ubiquitin binding and tubulin deacetylase) mark a distinct departure of HDAC6 from the known action of other HDACs. Recent discoveries have shown that HDAC6 is a key regulator of cytotoxic stress resistance (reviewed in Matthias et al. 2008). It appears to be both a sensor of stressful environmental stimuli and an effector, which mediates and coordinates appropriate cell responses.

Materials and Methods

Drosophila Lines and DNA Sequencing

DNA sequence data were collected from 12 highly inbred lines sampled in Africa (Lake Kariba, Zimbabwe). Furthermore, sequence data were obtained from 12 inbred

Key words: *Drosophila melanogaster*, selective sweep, selection mapping, *HDAC6*.

E-mail: svetec@bio.lmu.de.

Mol. Biol. Evol. 26(7):1549–1556. 2009
doi:10.1093/molbev/msp065
Advance Access publication April 6, 2009

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

European lines from The Netherlands. Both samples are described in detail in Glinka et al. (2003). All *Drosophila* strains were kept at 23 °C in glass bottles of 250 ml containing 80 ml standard cornmeal and yeast medium under a 6–18 dark–light cycle with 45% humidity.

DNA primers were designed based on the *D. melanogaster* genome sequence (http://flybase.org/) and obtained from Metabion (Martinsried, Germany). Genomic DNA from each line was extracted from pools of 20 females using the Puregene DNA isolation kit (Gentra System, Minneapolis, MN). Short DNA fragments of about 300–700 bp long were amplified by standard polymerase chain reaction (PCR) using the Taq DNA polymerase recombinant kit (Invitrogen, Carlsbad, CA). PCR products were purified using the Exosap-It kit (USB, Cleveland), and sequence reactions were conducted with ABI PRISM Big Dye Terminator v1.1. Sequence data were then obtained by an ABI 3730 DNA analyzer (Applied Biosystems + Hitachi, Foster City, CA).

Sequence edition and alignments were performed with the DNAstar software package, including Editseq, Seqman, and Megalign (DNASTAR, Madison, WI). Alignments were performed using the ClustalV option of Megalign. However, in cases of ambiguous alignments, we manually chose the most parsimonious scenario. Insertion and deletion polymorphisms were excluded from further analysis. Absolute positions of the DNA sequence follow the Flybase release 5.10.

Mapping Strategy

To identify and map the target of selection, we proceeded as follows: First, we selected a subgenomic region of about 70 kb on the X chromosome that contained several ecologically interesting genes, including a gene encoding a putative antifreeze protein (*CG6227*). This region partially overlaps with the window 47 in Li and Stephan (2006). Resequencing an additional (limited) number of short fragments of 500–600 bp in the 70-kb subgenomic region, we found very low levels of variation across most of the region in the European sample (data not shown), whereas the valley of reduced variation in the African sample appeared much narrower, that is, the situation was similar as in the case of the "*roughest*" and "*wapl*" regions (Beisswanger et al. 2006; Pool et al. 2006). To be able to localize the target of selection as precisely as possible, we therefore decided to follow the same strategy as in the *wapl* analysis (Beisswanger and Stephan 2008) and concentrated on the African sample (see "Standard analyses of a candidate region of selection" in the Results section). In a second step, we narrowed this 70-kb region down to 22 kb, resequenced this segment completely, and applied the specific tests for selective sweeps to this region (see Results).

Outlier Analysis

We used DnaSP 4.50.3 (Rozas et al. 2003) to calculate the basic summary statistics π, θ_W, Tajima's *D* (Tajima 1989), divergence, Fu and Li's *D* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000). Divergence was calculated between the sample from the African population of *D. melanogaster* and the available online release of the *Drosophila simulans* sequence (Flybase consortium; http://www.flybase.org). The ancestral states were defined using either *D. simulans* or (when not available) its close relative *Drosophila sechellia*.

We compared the mean value of each summary statistic of the 70-kb candidate region with its average value obtained for the whole X chromosome (Ometto et al. 2005). For each summary statistic, we used the Mann–Whitney test to infer whether the region represents an outlier compared with the rest of the X chromosome.

Ascertainment Bias Correction

Thornton and Jensen (2007) describe an approach that generates a uniform distribution of *P* values when some of the assumptions of the neutrality tests are violated. They study cases when past demographic events have shaped the polymorphism patterns of a subgenomic region, which is a biased sample based on a priori information (e.g., from a genome scan). The *HDAC6* subgenomic region was selected based on the genes in this region that may contribute to the ecological adaptation of *D. melanogaster*. Even if such a sampling is not random, it is unclear whether it generates any bias on selective sweep scanning and how to sample conditional on this biological information.

Performing a genome scan analysis, Li and Stephan (2006) discovered a 100-kb fragment that overlaps with the *HDAC6* region and showed evidence of recent positive selection in the European population of *D. melanogaster*. Among the fragments Li and Stephan (2006) analyzed was a 560-bp fragment located within the *HDAC6* subgenomic region that contained no polymorphic sites. This information was not considered important for the initial choice of the 70-kb region. However, we decided to include it in the analysis as a priori information making this analysis more conservative. Thus, we simulate a sample of 24 lines (12 European and 12 African ones) according to the demographic scenario inferred by Li and Stephan (2006). Conditioning on the existence of a monomorphic 560-bp fragment within the European sample, we create the null distribution of the neutrality test statistics used in this paper.

Composite Likelihood Ratio (CLR) Test

The CLR test (Kim and Stephan 2002) was used to infer selection. It computes the CLR (Λ_{CLR}) between a standard neutral model and a selective sweep model. The null distribution of the statistic is derived using the approach described in the "Ascertainment bias correction" section (see also fig. 1). This modification follows a suggestion of Thornton and Jensen (2007) who showed that the false positive rate can be controlled if the correct demographic null model is used. For the generation of the simulated data sets, we used the estimated value of the parameter θ_W (0.0499) under the demographic scenario of figure 1. Furthermore, the B test of the Kim and Stephan (2002) method was performed because it is more conservative. The CLR test was also used to estimate the target site of selection. However,

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

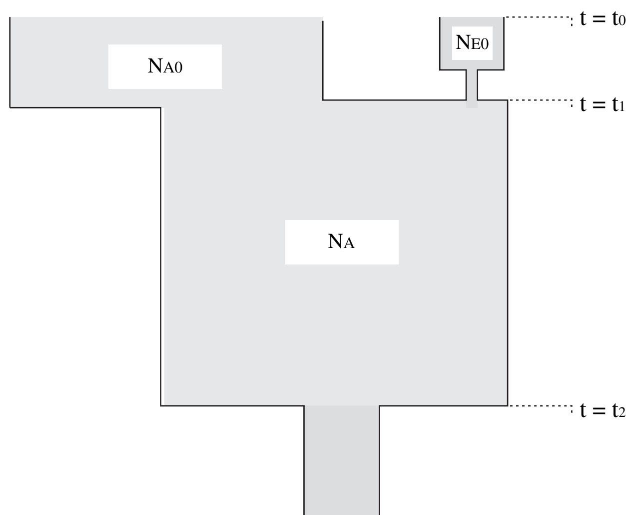

FIG. 1.—The demographic model of the European and African population of *Drosophila melanogaster* as it was inferred by Li and Stephan (2006) and used in this study. The present European effective population size is approximately $N_{E0} = 10^6$, whereas the African population (N_{A0}) is eight times larger. Backward in time, the model can be described by a severe bottleneck in the European population that took place $t_1 = 15{,}460$ years ago and lasted for ~340 years. During the bottleneck, the effective population size of the European population was decreased to 2,200. Approximately at $t_2 = 15{,}800$ years ago, the European population merges with the African population forming the ancestral population ($N_A = N_{A0}$). Finally, the ancestral population decreases to a fifth of the present day African population at $t_3 = 60{,}000$ years ago.

its confidence interval (CI) could not be determined (in contrast to Beisswanger and Stephan 2008), as population recombination rate was too high to run simulations of the sweep model in reasonable times.

"SweepFinder" Test

To infer selection, we also used the SweepFinder test. It takes into account the SFS of the whole chromosome (background SFS) in order to calculate the likelihood of the neutral model. Nonpolymorphic sites were excluded from the analysis, as Nielsen et al. (2005) suggest. SweepFinder uses the same principles as the CLR test: By comparing two hypotheses, a model of neutral evolution and a model of a selective sweep that just completed, it calculates the maximum likelihood estimates of the position of the beneficial allele as well as the strength of selection. Additionally, it reports the likelihood ratio Λ_{SF} between the null and the alternative model. Similarly to the CLR test, a null distribution is required to decide about the statistical significance of the selective sweep hypothesis. The main advantage of the SweepFinder is that a specific population genetic model is not considered in the null hypothesis, but the SFS is derived from the whole-chromosomal pattern of variation, that is, from the data itself.

We have extended the original approach for calculating the significance threshold for the SweepFinder. According to Nielsen et al. (2005), the 95th percentile of the statistic Λ_{SF} denotes the threshold value. Our approach, however, splits the region of interest into k fragments and for each one the $100 - (5/k)$ percentile is used as the cutoff value, resulting in a variable region-specific threshold. This approach helps to remedy the tendency of the SweepFinder to produce higher Λ_{SF} values at the borders of the region under study (Pavlidis P, unpublished results). Here, we chose $k = 10$. The demographic model of figure 1 (Li and Stephan 2006) with the ascertainment bias described in the "Ascertainment bias correction" section is used to create the null distribution of the test statistics for all performed neutrality tests.

Estimation of the Time Since Fixation of the Beneficial Allele

The time since the fixation of the beneficial allele was estimated by the methods described in Przeworski (2003) and Slatkin and Hudson (1991). For the Przeworski test, mutation rate $\mu = 1.45 \times 10^{-9}$/bp/gen (Li and Stephan 2006) and recombination rate $r = 4.718 \times 10^{-8}$/bp/gen (Comeron et al. 1999) were used. The local parameters were estimated from a 925-bp long region located between the seventh and ninth exons of *HDAC6* (as exon 8 is very short [88bp], it has presumably no special effect on the parameter estimates, and was thus kept in the analysis). This region contains 10 segregating sites forming eight haplotypes, and Tajima's $D = -1.74221$. Two positions of the beneficial mutation were tested: one in the last exon of *HDAC6* and one in the last exon of *CG9123*.

We also used the Slatkin–Hudson method (Slatkin and Hudson 1991) assuming a starlike genealogy since the fixation of the beneficial allele. We based this estimation on the DNA region between positions 9.865 and 12.443 kb. In this region, 19 segregating sites were detected, and divergence to *D. simulans* is 0.056. To convert the obtained estimates into years, we assumed 10 generations per year for both methods.

Results

Standard Analyses of a Candidate Region of Selection

The region analyzed here is about 70 kb long. It is located in a highly recombining portion of the X chromosome ($r = 4.718 \times 10^{-8}$/bp/gen) and is relatively gene dense. This region contains 12 genes, five of which have unknown molecular functions (*CG15032*, *CG9114*, *CG9123*, *CG12608*, and *CG9164*). The other genes have been functionally characterized (*gce*, *Top1*, *dah*, *HDAC6*, *CG6227*, *acj6*, and *Pp1*). In order to perform a fine-scale analysis of the African sample, we sequenced 15 noncoding (intronic or intergenic) DNA fragments of 511 bp on average, in addition to the four already sequenced by Ometto et al. (2005) (fig. 2). For each of these 19 fragments, basic summary statistics were calculated, averaged over the whole candidate region, and then compared with the chromosomal average. Only 15 of the 19 fragments could be aligned with *D. simulans*.

The region exhibits a strong reduction in nucleotide polymorphism. On average, the 259 fragments sequenced by Ometto et al. (2005) for the African population contained twice as many segregating sites as the 70-kb

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

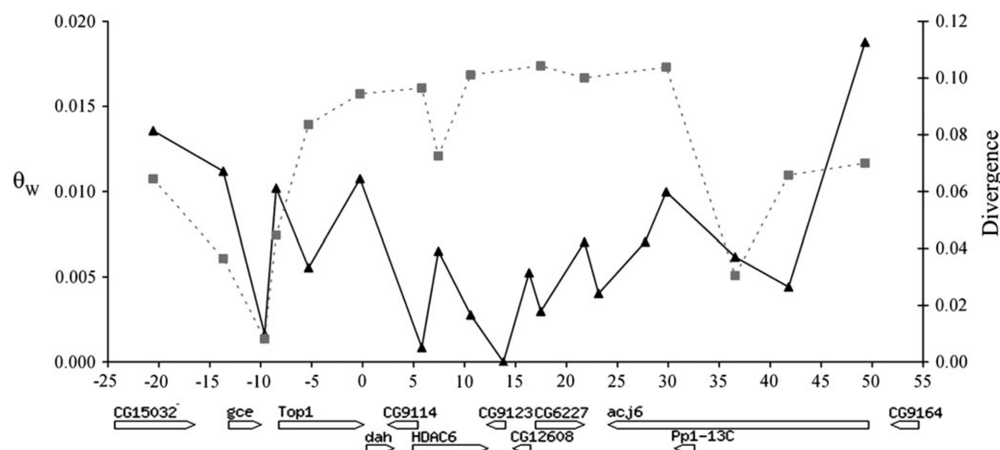

FIG. 2.—Nucleotide diversity θ_W (solid line) and divergence to *Drosophila simulans* (dashed line) across the candidate region for selection. The relative positions in kb are on the *X*-axis. Gene spans (according to Flybase) are at the bottom of the graph.

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

candidate region ($P < 0.0001$; see table 1). π and θ_W were significantly lower than the chromosomal average ($P < 0.0001$ for both). As can be seen in figure 2, the θ_W curve is roughly U shaped (with a minimum between 10 and 15 kb), except for two positions at −10 and around 40 kb where divergence is very low. In general, divergence is rather high in the region of reduced variation between positions 0 and 22 kb (~0.09).

Furthermore, the region shows deviations from the chromosomal expectation with regard to the SFS. Indeed, Tajima's *D* values are more negative than the X chromosome average (−1.143 vs. −0.667), which is highly significant ($P = 0.001$). Four fragments show significantly negative Tajima's *D* values (data not shown). In contrast, Fay and Wu's *H* statistic does not depart from the chromosomal average. This illustrates that the SFS is lacking intermediate frequency variants and shows an excess of low frequency single nucleotide polymorphisms (SNPs).

The number of haplotypes ranges from 1 to 12 in the candidate region, but its mean is significantly lower than the chromosomal average ($P < 0.001$). Similarly, haplotype diversity is significantly lower ($P < 0.001$). LD as measured by the *ZnS* statistic is relatively constant over the whole region (<0.3) and does not deviate from the chromosomal average.

The genes *CG9123* and *CG12608* are paralogs. Among the 12 *Drosophila* genomes examined (*Drosophila* 12 Genomes Consortium 2007), this duplication is present only in *D. melanogaster*. Both copies are highly diverged from *D. simulans*. Investigating the pattern of polymorphism at both genes, we did not find evidence for extensive gene conversion; for instance, there is only one SNP shared between both copies (of 48 SNPs in total). *CG9123* contains many nonsynonymous SNPs in relatively high frequency, most of which produce drastic amino acid changes (see supplementary table 1, Supplementary Material online). In addition, we observed some deletions in the coding region, one of which causes a frameshift change. This may suggest that *CG9123* is under weak functional constraints or even a pseudogene.

Table 1
Mann–Whitney Comparisons of Different Summary Statistics between the Candidate Region and the Chromosomal Mean

Summary Statistics	Chromosomal Mean	Candidate Region Mean	*P* Value
Recombination rate	3.515	4.714	<0.0001
Sample size	11	11	0.81
Fragment length	501	511	0.63
Segregating sites	19	11	<0.0001
θ_W	0.013	0.007	<0.0001
Π	0.012	0.005	<0.0001
Tajima's *D*	−0.667	−1.143	0.0001
Number of haplotypes	9.655	7.348	0.0001
Haplotype diversity	0.93	0.785	0.0001
ZnS	0.14	0.12	0.44
Divergence	0.064	0.068	0.36
Fay and Wu's *H*	−0.26	−0.32	0.56

Application of the CLR and SweepFinder Tests

In order to perform more advanced neutrality tests, we defined a region of about 22 kb (corresponding to the segment between absolute positions 15,222,319 and 15,244,496 in Flybase release 5.10, and to positions 0 to 22 kb in fig. 2). This region was then completely sequenced and subjected to the CLR and SweepFinder tests. The CLR test was marginally significant ($P = 0.048$) when the null distribution of the statistic Λ_{CLR} was constructed from the demographic scenario of the African population inferred by Li and Stephan (2006) (fig. 1). Figure 3*A* shows Λ_{CLR} along the region. The beneficial mutation is estimated to have occurred at position 11.378 kb relative to the beginning of the 22-kb region, and $\alpha = 2Ns$ is approximately 13,076 (where *N* is the effective population size and *s* the selection coefficient). This value is much higher than most other reported estimates, which is consistent with the observed width of the valley of reduced variation and the fact that population recombination rate $4Nr$ is very high in this part of the genome.

The SweepFinder test was also significant ($P = 0.034$) for the 22-kb completely sequenced region. In figure 3*B* we show the Λ_{SF} values along the region. Consistent with the

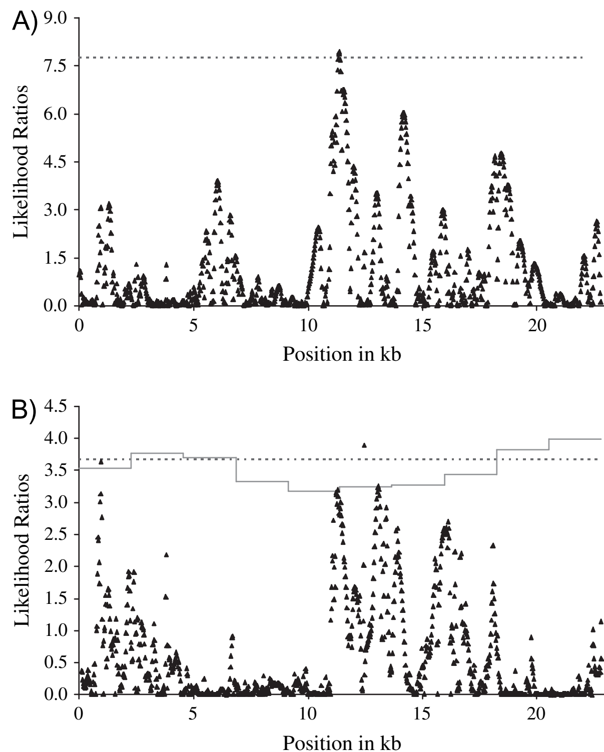

FIG. 3.—The likelihood-ratio values calculated by (*A*) the CLR and (*B*) the SweepFinder tests for a 22-kb subregion of the 70-kb region (for 1,000 bins). Each triangle denotes the value of the test statistics for a selective sweep model for which the beneficial mutation occurred at that specific position. In (*B*) the dashed line depicts the constant threshold calculated according to Nielsen et al. (2005), whereas the solid line shows the variable threshold (see Materials and Methods).

result of the CLR test, three positions (11.315, 12.474, and 13.110 kb) show the highest Λ_{SF} values. The high value around position 1.0 kb is probably not a target of selection as it is not confirmed by Tajima's *D* and the CLR test.

Age of the Selective Sweep

The age of the sweep in the 22-kb region was estimated by the Przeworski and Slatkin–Hudson methods (cf. Materials and Methods). We used Przeworski's approach with two positions as input parameter values that are near the estimated selected sites: Position 11.787 kb gave a time since fixation of the beneficial allele of 63,334 years (95% CI: 23,382–628,432 years), whereas position 12.787 kb gave 56,770 years (95% CI: 21,121–577,307 years). Using the Slatkin–Hudson method, the age of the sweep was estimated as 50,047 years.

These estimates suggest that the sweep occurred before the European lineage split off from the African one (about 16,000 years ago; Li and Stephan 2006). In order to confirm this hypothesis, we resequenced the region between position 8.0 and 15.0 kb in 12 lines of a European sample from The Netherlands (Materials and Methods). We found that the European lines were identical with those of the African sample in a limited segment of approximately 2.7 kb from position 9.8 to 12.5 kb (except for three derived singletons and one doubleton; see supplementary table 1, Supplementary Material online). This suggests, in conjunction with the estimated age of the sweep, that the selected allele has been exported to Europe during the colonization process.

Sliding Window Analysis

To corroborate our mapping results, we also performed a sliding window analysis on the SFS of the sequenced 22-kb region (fig. 4). Indeed, significantly negative Tajima's *D* and Fu and Li's *D* values were found near the estimated targets of selection, consistent with the CLR and SweepFinder results. The analysis revealed a small segment showing a local reduction of nucleotide diversity and an SFS shifted toward low-frequency variants despite normal levels of divergence. This region coincided with the 2.7-kb fragment mentioned above. Four exons lie in this region: the three last exons of *HDAC6* and a portion of the last exon of *CG9123*. The relatively low value of Tajima's *D* around position 19 kb is probably due to purifying selection (causing the observed low divergence in the helicase functional domain of *CG6227*; data not shown).

In order to identify candidate substitutions under selection, we aligned the 2.7-kb region of *D. melanogaster* to that of *Drosophila sechellia*, *D. simulans*, *Drosophila erecta*, and *Drosophila yakuba*. As the 2.7-kb region centers on *HDAC6*, we focused our investigations on this gene. The *HDAC6* introns were poorly conserved between species, but we obtained a good alignment of the 3′ untranslated region (UTR) and of the three last exons of the gene. In the 3′ UTR, we found six nucleotide substitutions specific to the *D. melanogaster* lineage. In exon 7, we identified three nonsynonymous substitutions specific to *D. melanogaster*. All of them cause nonpolar to nonpolar amino acid replacements. We also found a deletion of nine nucleotides that is specific to *D. melanogaster* at the end of exon 9 (see supplementary table 2, Supplementary Material online). This exon also carries two nonsynonymous substitutions. One of them generates a drastic amino acid change: a valine to glutamic acid substitution (see supplementary table 2, Supplementary Material online). In addition, this substitution is in a region predicted by the program MyHits (http://myhits.isb-sib.ch) to be the ubiquitin-binding site of HDAC6.

Discussion

Evidence for a Selective Sweep in the *HDAC6* Region of African *D. melanogaster*

By completely resequencing a 22-kb region around *HDAC6* in a sample of 12 African *D. melanogaster* X chromosomes and applying two likelihood tests (CLR and SweepFinder), we found evidence consistent with the presence of a selective sweep in this region. Furthermore, our mapping showed that the target of selection is most likely located in a 2.7-kb DNA region, centering on the last exon of *HDAC6*.

The expected age of the sweep was estimated as 50,000–63,000 years, depending on the method and input parameter values. This suggests that the sweep occurred before the European lineage split off from the African one

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

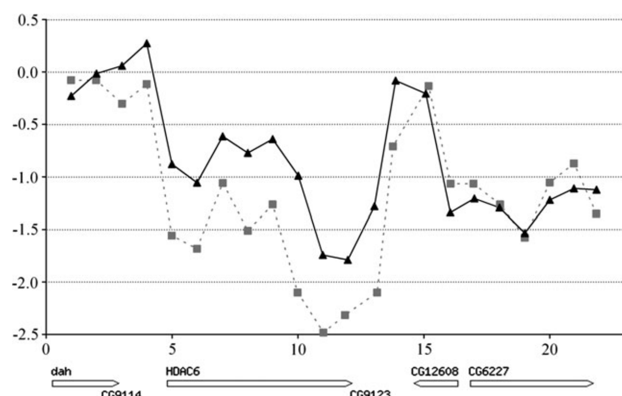

FIG. 4.—Sliding window analysis of the fully sequenced 22-kb region. Tajima's *D* and Fu and Li's *D* are represented by solid lines (black triangles) and dashed lines (gray squares), respectively. Each data point represents the midpoint of a 2,000-bp long window and the step size is 1,000 bp. In order to describe the neutral frequency spectrum, we excluded the nonsynonymous sites from this analysis.

(which occurred about 16,000 years ago; Li and Stephan 2006). Our age estimates are consistent with the observation that the sequences of the *HDAC6* alleles from our European sample are identical with that of the African haplotype in the swept region of approximately 2.7 kb (except for some derived low-frequency variants). Consistent with the relatively old age of the selective sweep, we did not identify any pattern of LD that is characteristic of a sweep (according to the predictions by Pfaffelhuber et al. 2008). Interestingly, a PAML analysis (Yang 2007) of *HDAC6* sequences from five species of the *D. melanogaster* subgroup found no evidence of selection (data not shown). This suggests that, prior to the inferred selective sweep, *HDAC6* has not undergone accelerated evolution in the past few million years.

It is clear that the evidence we provided is subject to some uncertainty. First, the results inferred by the CLR and SweepFinder tests may depend to some degree on demography. In particular, complex demographies could be a confounding factor (for instance, population size bottlenecks; Pavlidis et al. 2008). However, the demographic history of the African population we inferred previously is probably relatively simple and may be summarized by an expansion model (Li and Stephan 2006; Hutter et al. 2007). Furthermore, we have improved the original CLR test by Kim and Stephan (2002) and have now taken demography into account. Finally, the problem of demography is alleviated by applying SweepFinder, because the chromosomewide background SFS is used rather than a specific model. Second, a more general concern may be that if selection is a frequent and major pervasive force our two-step approach for inferring selection may not work (Hahn 2008). Then, a joint inference of selective and demographic parameters would be a more appropriate approach. However, we emphasize that we search for very strong selection. In such a case, our method of separating demography from selection is expected to be a reasonable first approximation. Third, the uncertainty in the estimates of the target site of selection needs to be mentioned. Unlike Beisswanger and Stephan (2008), we were not able to obtain confidence intervals of our estimates, as the rate of recombination in the *HDAC6* region is too large. However, based on the SFS, we were able to support our conclusion that the most likely target of selection is located in a 2.7-kb region (between positions 9.8 and 12.5 kb; see fig. 4). This result is consistent with the observation that the European alleles are identical in this region with the selected African allele. This latter argument, however, requires that the sweep occurred in Africa before the African and European lineages split, which is indeed supported by the estimated lower bound of the age of the sweep of >20,000 years.

Can the polymorphism patterns in the *HDAC6* region be explained by selective pressures other than positive directional selection? It is possible that at least part of the polymorphism pattern is associated with the action of purifying selection. The entire 70-kb region contains several functional elements that give rise to low divergence levels (fig. 2). In the identified 2.7-kb region between positions 9.8 and 12.5 kb, however, divergence is everywhere in the range of 5–8% and thus comparable with the average of the whole 70-kb region of 6.8%. This suggests that purifying selection is not likely a major cause of the observed pattern of variation in the 2.7-kb region.

Significance of the Selective Sweep in Relation to the Function of *HDAC6*

The 2.7-kb region we mapped by the selection approach overlaps with the last exons of two genes, *HDAC6* and *CG9123*. The latter is a duplicate of *CG12608*. According to the alignment of the 12 fully sequenced *Drosophila* species (*Drosophila* 12 Genomes Consortium 2007), this duplication event occurred in the *D. melanogaster* lineage. However, based on the polymorphism pattern mentioned above, *CG9123* is probably a pseudogene (or on its way to becoming one). Furthermore, *CG9123* is located at the boundary of the identified 2.7-kb region. We therefore concentrate the following discussion on *HDAC6*.

HDAC6 is a unique member of the histone deacetylase family harboring a ubiquitin-binding site and two catalytic deacetylase domains (Verdel et al. 2000; Khochbin et al. 2001). In addition, its localization in the cytoplasm is very unusual for a histone deacetylase (Verdel et al. 2000). It has been shown that its role is not limited to gene regulation. Rather, it is also important for the general cytotoxic stress response. It is involved in the two major cellular mechanisms degrading misfolded protein aggregates: autophagy and the ubiquitin–proteasome system (Pandey et al. 2007). HDAC6 detects and mediates the cytotoxic stress response at three different levels: First, its strong ubiquitin-binding ability coupled with its ability to move along microtubules allows HDAC6 to transport ubiquitinated protein aggregates, thus favoring the formation of aggresomes. Second, HDAC6 is able to stimulate autophagy when the ubiquitin–proteasome system is impaired (Pandey et al. 2007), and finally it mediates the activation of heat shock proteins (Boyault et al. 2006). More generally, HDAC6 is believed to be involved in several other cell stress response pathways such as antiviral responses (Boyault et al. 2006).

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

In *D. melanogaster*, *HDAC6* is mainly expressed in an insect specific organ: the Malpighian tubule (Chintapalli et al. 2007). Its tissues might be exposed to a broad range of cellular stress as it carries out most of the osmoregulation and the excretion of organic solutes as well as xenobiotics (Dow and Davies 2006).

To identify possible targets of selection, we aligned the *HDAC6* sequence of five *Drosophila* species. It revealed that HDAC6 carries a limited number of *D. melanogaster*-specific changes. But we could neither confirm nor exclude that any of them is a positively selected substitution. Indeed, any nucleotide change in the introns or 3′ UTR could affect *HDAC6*'s regulation or expression and any of the nonsynonymous changes observed in the exons could modify the protein's properties. However, in the last exon of *HDAC6*, one nonsynonymous substitution may well have significant functional consequences: a valine-to-glutamic acid replacement that occurred in the *D. melanogaster* lineage and is located in the ubiquitin-binding site of *HDAC6*. Could this substitution affect the ubiquitin-binding affinity of HDAC6 and thus the response of cells to stress? Ubiquitin-binding assays (Boyault et al. 2006) comparing the *D. melanogaster* and *D. simulans* alleles may provide an answer to this question.

Supplementary Material

Supplementary tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Sequences are available in the GenBank under the accession numbers FJ764835–FJ764973.

Acknowledgments

We thank the Munich *Drosophila* group for valuable discussions, Stefan Laurent for performing the PAML analysis, and Simone Lange and Anne Wilken for excellent technical assistance. Furthermore, we thank two anonymous referees for detailed comments that improved the manuscript. This research was supported by grants from the Deutsche Forschungsgemeinschaft to W.S. (grant Ste 325/7 and grant 325/12 of the Research Unit 1078) and the VolkswagenStiftung to P.P. (I/824234).

Literature Cited

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2:e286.

Beisswanger S, Stephan W. 2008. Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. Proc Natl Acad Sci USA. 105:5447–5452.

Beisswanger S, Stephan W, De Lorenzo D. 2006. Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. Genetics. 172:265–274.

Boyault C, Gilquin B, Zhang Y, Rybin V, Garman E, Meyer-Klaucke W, Matthias P, Muller CW, Khochbin S. 2006. HDAC6-p97/VCP controlled polyubiquitin chain turnover. Embo J. 25:3357–3366.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 140:783–796.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet. 39:715–720.

Comeron JM, Kreitman M, Aguadé M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics. 151:239–249.

Dow JA, Davies SA. 2006. The Malpighian tubule: rapid insights from post-genomic biology. J Insect Physiol. 52:365–378.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450: 203–218.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–1413.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics. 165:1269–1278.

Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution. 62:255–265.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics. 177:469–480.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature. 449:851–861.

Jensen JD, Bauer DuMont VL, Ashmore AB, Gutierrez A, Aquadro CF. 2007. Patterns of sequence variability and divergence at the *diminutive* gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. Genetics. 177:1071–1085.

Jensen JD, Kim Y, Bauer DuMont VL, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 170:1401–1410.

Khochbin S, Verdel A, Lemercier C, Seigneurin-Berny D. 2001. Functional significance of histone deacetylase diversity. Curr Opin Genet Dev. 11:162–166.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics. 167:1513–1524.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 160:765–777.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2:e166.

Matthias P, Yoshida M, Khochbin S. 2008. HDAC6 a new cellular stress surveillance factor. Cell Cycle. 7:7–10.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–35.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 22:2119–2130.

Orengo DJ, Aguadé M. 2004. Detecting the footprint of positive selection in a European population of *Drosophila*

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015

melanogaster: multilocus pattern of variation and distance to coding regions. Genetics. 167:1759–1766.

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.

Pandey UB, Nie Z, Batlevi Y, McCray BA, et al. (18 co-authors). 2007. HDAC6 rescues neurodegeneration and provides an essential link between autophagy and the UPS. Nature. 447:859–863.

Pavlidis P, Hutter S, Stephan W. 2008. A population genomic approach to map recent positive selection in model species. Mol Ecol. 17:3585–3598.

Pfaffelhuber P, Lehnert A, Stephan W. 2008. Linkage disequilibrium under genetic hitchhiking in finite populations. Genetics. 179:527–537.

Pool JE, Bauer DuMont VL, Mueller JL, Aquadro CF. 2006. A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. Genetics. 172:1093–1105.

Przeworski M. 2003. Estimating the time since the fixation of a beneficial allele. Genetics. 164:1667–1676.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 129:555–562.

Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 172:2647–2663.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123: 585–595.

Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 175:737–750.

Verdel A, Curtet S, Brocard MP, Rousseaux S, Lemercier C, Yoshida M, Khochbin S. 2000. Active maintenance of mHDA2/mHDAC6 histone-deacetylase in the cytoplasm. Curr Biol. 10:747–749.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet. 3:e90.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Rasmus Nielsen, Associate Editor

Accepted March 26, 2009

Downloaded from http://mbe.oxfordjournals.org/ at Forth Library on March 23, 2015