MBE Advance Access published November 18, 2011

Research article

Selective Sweep of a *cis*-Regulatory Sequence in a Non-African Population of *Drosophila melanogaster*

Sarah S. Saminadin-Peter,*[,1] Claus Kemkemer,* Pavlos Pavlidis,[2] and John Parsch

Department of Biology II, University of Munich (LMU), Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany

*These authors contributed equally to this work

[1]**Present address:** Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA

[2]**Present address:** Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Corresponding author:

John Parsch

Department Biology II, University of Munich (LMU)

Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany

Tel: +49 89 2180 74 107, Fax: +49 89 2180 74 104

Email: parsch@bio.lmu.de

Running head: *cis*-regulatory selective sweep

Key words: gene expression, adaptation, selection, population structure, regulatory evolution

Abstract

Although it is thought that changes in gene expression play an important role in adaptation, the identification of gene-regulatory sequences that have been targets of positive selection has proven difficult. Here we identify a *cis*-regulatory element of the *Drosophila melanogaster CG9509* gene that is associated with a selective sweep in a derived, non-African population of the species. Expression analyses indicate that *CG9509* consistently shows greater expression in non-African than in African strains of *D. melanogaster*. We find that a 1.8-kb region located just upstream of the *CG9509* coding region is devoid of DNA sequence polymorphism in a European population sample and that this is best explained by the recent action of positive selection (within the past 4,000–10,000 years). Using a reporter gene construct and phiC31-mediated site-specific integration, we show that the European version of the *CG9509* upstream region drives 2-3 times greater expression than the African version in an otherwise identical genetic background. This expression difference corresponds well to that of the native gene and indicates that sequence variation within the *CG9509* upstream region can completely account for its high expression in the European population. Selection appears to have favored a quantitative increase in gene expression in the Malphigian tubule, the tissue where *CG9509* is predominantly expressed.

Introduction

The identification of genetic changes that underlie adaptation has been a long-standing goal of evolutionary genetics. Such changes can be classified broadly into two classes: structural changes, which alter the amino acid sequence of a protein, and regulatory changes, which alter the amount and/or spatial distribution of gene expression. Examples of the former are numerous (Hoekstra and Coyne 2007) and statistical methods are available that allow one to infer the adaptive evolution of protein sequences from inter- and intra-species sequence comparisons (e.g., McDonald and Kreitman 1991; Fay et al. 2001; Smith and Eyre-Walker 2002; Bustamante et al. 2002; Sawyer et al. 2007). These methods have suggested that 45–95% of amino acid replacements between *Drosophila* species are the result of positive selection (Fay et al. 2002; Smith and Eyre-Walker 2002; Sawyer et al. 2007; Maside and Charlesworth 2007; Haddrill et al. 2008; Bachtrog 2008; Grath et al. 2009). Examples of adaptive regulatory changes are much more scarce (Hoekstra and Coyne 2007), however this may reflect the fact that statistical methods for detecting regulatory adaptation are lacking and the molecular characterization of putative gene-regulatory elements often requires extensive experimental work (e.g., Gompel et al. 2005; Prud'homme et al. 2006; McGregor et al. 2007; Jeong et al. 2008). Several authors have proposed that regulatory change plays an even greater role than structural change in the evolution of organismal morphology (King and Wilson 1975; Carroll 2000; Wray 2007; Carroll 2008).

Perhaps the best example of adaptive regulatory evolution in *Drosophila melanogaster* is the cytochrome P450 gene, *Cyp6g1*, where an *Accord* transposable element inserted just upstream of the *Cyp6g1* coding region has been shown to increase

Cyp6g1 expression and confer greater resistance to DDT and related insecticides (Daborn et al. 2002; Chung et al. 2007). The *Accord* insertion is present at high frequency in worldwide populations, but at lower frequency in populations from the ancestral species range in east Africa – a pattern thought to result from natural selection favoring higher *Cyp6g1* expression in regions subjected to insecticide use (Catania et al. 2004). Although *Cyp6g1* is a prime example of adaptation at the level of gene expression, its discovery was possible only because it was linked to a conspicuous phenotype (DDT resistance) that could be used to genetically map the underlying DNA sequence polymorphism (Daborn et al. 2002). Thus, the *Cyp6g1* polymorphism may represent only the tip of the iceberg in terms of adaptive regulatory changes, as most such changes probably do not have a conspicuous effect on morphology or physiology.

A microarray analysis of gene expression variation among adult males from two natural populations of *D. melanogaster*, one from the derived species range in Europe (the Netherlands) and one from the ancestral range in Africa (Zimbabwe) identified over 150 genes that differed significantly in their expression level between populations (Hutter et al. 2008). The gene showing the greatest over-expression in the European population was the insecticide resistance gene *Cyp6g1* mentioned above. The gene showing the second highest over-expression in Europe was *CG9509*, a gene whose precise function is unknown, but which shows highly enriched expression in Malphigian tubule (Chintapalli et al. 2007) and is predicted to be a choline dehydrogenase (Tweedie et al. 2009). Interestingly, a previous microarray analysis identified *CG9509* as one of only 12 genes that differed significantly in expression between four cosmopolitan laboratory strains (from North America and Asia) and four Zimbabwe strains of *D. melanogaster*

(Meiklejohn et al. 2003). Given the low level of gene expression polymorphism within populations and the tendency for gene expression levels to be under stabilizing selection (Denver et al. 2005; Rifkin et al. 2005; Lemos et al. 2005; Gilad et al. 2006; Hutter et al. 2008), the large and consistent between-population difference in *CG9509* expression makes it a strong candidate for a gene having undergone adaptive regulatory evolution in association with the out-of-Africa expansion of *D. melanogaster* that occurred about 15,000 years ago (David and Capy 1988; Haddrill et al. 2005; Ometto et al. 2005; Thornton and Andolfatto 2006; Stephan and Li 2007).

Here we use a combination of population genetic and functional analyses to show that a DNA sequence region immediately upstream of the *CG9509* gene has been the target of a selective sweep in the European population of *D. melanogaster*. Furthermore, we use transgenic reporter gene constructs to demonstrate that the *CG9509* expression difference observed between European and African strains can be explained completely by sequence variation within a 1.2-kb fragment of the selected region. These results are consistent with positive selection favoring a *cis*-regulatory polymorphism that confers a quantitative change in the expression level of a single gene between populations.

Materials and Methods

Fly strains

For the survey of DNA sequence polymorphism, we used 12 isofemale lines from Africa (Lake Kariba, Zimbabwe) and 12 isofemale lines from Europe (Leiden, The Netherlands). These same lines were used in previous studies of X-linked and autosomal sequence polymorphism (Glinka et al. 2003; Ometto et al. 2005; Hutter et al. 2007;

Parsch et al. 2009). For qRT-PCR analysis, we used a subset of eight strains from each population, corresponding to those used in a previous microarray study (Hutter et al. 2008). Two stocks with mapped *attP* sites on the third chromosome, *phiX-68E* and *phiX-86Fb* (Bischof et al. 2007), were obtained from the Bloomington Stock Center (Indiana, USA) and used for phiC31 site-specific integration.

Expression analysis

Total RNA was extracted from 45 adult males or 30 adult females (4-6 days of age) using Trizol reagent (Invitrogen, Carlsbad, CA, USA) and the manufacturer's protocol. Reverse transcription was carried out using 5 mg of total RNA, random hexamer primers, and Superscript II reverse transcriptase (Invitrogen). The resulting cDNA was used for qRT-PCR with a TaqMan probe specific to *CG9509* (Dm01838873_g1) and one specific to the endogenous control gene *RpL32* (Dm02151827_g1). The *CG9509* assay is specific to a 60 bp cDNA amplicon spanning the exon1/exon2 boundary of transcript *CG9509-RA*. The *RpL32* assay is specific to a 72 bp cDNA amplicon spanning the exon1/exon2 boundary of transcript *RpL32-RA*. The average threshold cycle (*Ct*) was determined over three replicates per strain and *ΔCt* was calculated as the mean difference in *Ct* between the *CG9509* and the *RpL32* probe for each strain. The fold-change difference in expression between the European and African populations was calculated as $2^{-(\Delta Ct1 - \Delta Ct2)}$, where *ΔCt1* and *ΔCt2* represent the mean *ΔCt* values of the European and African strains, respectively. Assays were performed with a 7500 Fast Real-Time PCR System (Applied Biosciences, Foster City, CA, USA).

DNA sequencing

From each strain, the region containing the *CG9509* gene was PCR-amplified from genomic DNA in six separate reactions using the following primer pairs (all sequences are 5' – 3'): GCCCCTGTTCAATTTATTCG and TTCTGAATCGGCATCATCAC, GGCTGCAGCTCTTAAATGGC and ACGAGGACGTTGACTTAGCC, CCAATGGCTAAGTCAACGTCC and CAAAGAATAGTGCCGGCAAC, CCCACACCAACACCATACC and CTCCACATATGGCTGTCCCAAC, GATGGTCGCTGCTATTGGC and CTTGAATGGATAGACCCTTGG, ACGCAATCTCCAGGATCATGTC and CGTGGGCTAAACTTGTTGCTAAG. Following PCR, the amplified products were purified with ExoSAP-IT (USB, Cleveland, OH) and sequenced from both strands using the above primers, BigDye chemistry, and a 3730 automated sequencer (Applied Biosystems).

Statistical analysis

Standard DNA polymorphism and divergence statistics, as well as McDonald-Kreitman tests (McDonald and Kreitman 1991), we calculated using DnaSP v5 (Librado and Rozas 2009). To test for selective sweeps, we used the composite likelihood ratio (CLR; Kim and Stephan 2002), *SweepFinder* (Nielsen et al. 2005), and 'distance' tests. The last of these uses the maximum distance (in bp) between two polymorphic sites in a sample of alleles as the test statistic. In all cases, we calculated the *P*-value of the test statistic using a parametric bootstrapping approach that approximated the distribution of the test statistic from 1,000 simulated data sets. Simulations were performed using *ms* (Hudson 2002), assuming either a standard neutral model or a bottleneck model with parameters taken

from Li and Stephan (2006). In addition, we used machine learning and Bayesian methods (Pavlidis et al. 2010; Csilléry et al. 2010) to test for selective sweeps and to infer parameters, such as the selection coefficient (*s*) and the time since the sweep began (*t*). A complete description of these methods is provided in Supplementary Methods (Supplementary Material online).

Reporter gene assays

The region upstream of *CG9509* was PCR-amplified from one European and one African strain using the primers 5'-TGGCGCTAACCTGAATTCC-3' and 5'-GCGTTTTGCTTTTCCGTTAG-3'. The amplified region begins 9 bp after the *CG14406* stop codon and ends 2 bp before the *CG9509* start codon. The PCR products were cloned directly into the p*CR2.1-TOPO* vector (Invitrogen), with their identity and orientation being confirmed by restriction analysis. A 3.6-kb *Not*I fragment of the p*CMV-SPORT-βgal* plasmid (Invitrogen) containing the *E. coli lacZ* coding region was then inserted into the *Not*I site of the above plasmids and restriction analysis was performed to ensure that both the promoter and *lacZ* coding sequences were in the same orientation. As a final step, *BamH*I/*Xba*I fragments containing the promoter and the *lacZ* coding sequences were ligated into the p*attB* integration vector (Bischof et al. 2007).

Integration vector DNA was purified with the QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany) and eluted from the column with injection buffer (0.1 mM sodium phosphate pH 6.8, 5 mM KCl). Vector DNA at a concentration of 200 ng/μl was used for microinjection of early-stage embryos of the *phiX-68E* (*attP* site at cytological band 68E) and *phiX-86Fb* (*attP* site at cytological band 86F) strains, both of which

contain a stable source of phiC31 integrase on the X chromosome. After microinjection, surviving flies were crossed to a *white*$^-$ strain to remove the integrase source and establish stable lines. The offspring of this cross were screened for red eye color (imparted by the wild-type *white* gene of the vector), which was diagnostic for stable transformants.

Reporter gene expression was measured at the protein level by a β-galactosidase assay (Hense et al. 2007). For this, we used batches of five whole males or females (four-to-six days old) or dissected Malphigian tubules from 10 four-to-six day-old males. A minimum of two biological replicates, each with two technical replicates, were performed for each sex, integration site, and tissue. To measure *lacZ* transcript abundance, we performed qRT-PCR as described above using a custom TaqMan probe for the *lacZ* gene (Applied Biosystems; forward primer: 5'-GCTGGGATCTGCCATTGTCA-3', reverse primer: 5'CAGCGCAGACCGTTTTCG-3', FAM-labeled primer: 5'-CCCCGTACGTCTTCC-3') and the *RpL32* probe (Dm 02151827_g1) as an endogenous control. Two biological replicates, each with two technical replicates, were performed for each sex and integration site.

Results

Expression divergence of *CG9509* between African and non-African populations

Two previous microarray studies that examined adult male gene expression in *D. melanogaster* identified *CG9509* as showing a large and highly significant expression difference between African (Zimbabwe) and non-African (North American, Asian, or European) strains (Meiklejohn et al. 2003; Hutter et al. 2008). In both cases, the non-African strains had >2-fold higher expression than the African strains (table 1). To

confirm the observed expression difference in males and test whether it also extends to females, we performed quantitative reverse-transcription PCR (qRT-PCR) assays on flies of both sexes of the European strains and Zimbabwe strains used previously (Hutter et al. 2008). In both males and females, there was significantly higher *CG9509* expression in the European strains, with the magnitude of the expression difference agreeing well with that measured in the microarray experiments (table 1). In all of the above experiments, the flies had been reared for many generations in a common laboratory environment, indicating that the expression difference has a genetic, rather than an environmental basis.

DNA sequence polymorphism in the *CG9509* region

To determine if the expression difference of *CG9509* was associated with genetic variation linked to the gene itself (i.e., *cis*-regulatory polymorphism), we sequenced a 5.6-kb region of the X chromosome encompassing the entire *CG9509* coding region and ~3 kb of its 5' flanking region in 12 strains each of the European and Zimbabwe populations, including the strains used in the expression analyses. The *CG9509* upstream region includes the complete transcriptional unit of the gene *CG14406* and the 3' end of the *CG12398* transcriptional unit (fig. 1). Across the whole region, the per-nucleotide estimate of sequence diversity, θ, in the European population was 0.0021, which is lower than the mean value (0.0047) for the X chromosome in this population (Hutter et al. 2007). Particularly striking was a 1.8-kb region just upstream of the *CG9509* coding region that is devoid of nucleotide polymorphism in the European population. Across this region, only a single haplotype is present in the European sample. In contrast, this region shows normal levels of polymorphism within the African population, as well as normal

levels of divergence between *D. melanogaster* and *D. sechellia* (fig. 1). Thus, we can rule out the possibility that the lack of polymorphism in the European population is the result of this region being under abnormally high selective constraint or having an unusually low mutation rate. A comparison of polymorphism within *D. melanogaster* to divergence with *D. sechellia* provided evidence for recurrent positive selection acting on both the coding and upstream regions of *CG9509* by the test of McDonald and Kreitman (1991) (table 2).

Tests for selective sweeps

The absence of nucleotide polymorphism in the *CG9509* upstream region suggests that there may have been a selective sweep of a beneficial allele in the European population. We tested this possibility using the CLR test of Kim and Stephan (2002), the *SweepFinder* test of Nielsen et al. (2005), and a 'distance' test, which examines the greatest distance (in bp) between two polymorphic sites in a sample of alleles. All three methods rejected a null hypothesis of standard, neutral evolution (e.g., no selection, constant population size) with high confidence ($P < 0.001$ in all cases). However, when the null hypothesis included a population bottleneck with parameters relevant to the European *D. melanogaster* population (Li and Stephan 2006), only the distance test remained significant ($P = 0.028$), while the CLR and *SweepFinder* tests were marginally significant ($P = 0.123$ and $P = 0.068$, respectively).

To better distinguish between selection and demography as the cause of reduced polymorphism in the European population, we performed coalescent simulations of nucleotide sequence polymorphism data both with and without selection (Ewing and

Hermisson 2010) using demographic parameters relevant to our populations (Li and Stephan 2006). The simulated data were used in a machine learning approach (Pavlidis et al. 2010) to classify neutral and selected patterns of polymorphism on the basis of two statistics, *SweepFinder* (Nielsen et al. 2005) and ω (Kim and Nielsen 2004), called features. Our observed data fell into the selected range (fig. 2*A*), although the false positive rate over the entire feature space was high (28%). This is expected, because bottleneck models with selection generate similar polymorphism patterns as neutral bottleneck models (Pavlidis et al. 2010). However, in contrast to the high false positive rate in the global simulation space, a local analysis revealed a clear over-representation of selection instances around the observed data point (fig. 2*B*). The probability of observing such a high proportion of selected data points in the local area around the observed data point by chance is very low ($\sim 10^{-8}$), indicating a significant over-representation of selection instances in this region of the feature space.

We also used Approximate Bayesian Computation (ABC) to estimate the selection coefficient (s) associated with the selective sweep and the number of generations (t) since the sweep began. The posterior distributions of these parameters (fig. 3) are consistent with strong and recent positive selection acting within the *CG9509* genomic region. For s, the mode, mean and median of the posterior distribution are 0.089, 0.061, and 0.065, respectively. Assuming an effective population size (N_e) of $\sim 10^6$ (Li and Stephan 2006), this corresponds to a scaled selection coefficient ($N_e s$) of ~60,000. For t, the mode, mean and median of the posterior distribution are 4,871, 39,875, and 20,899, respectively. Assuming 10 generations per year, this corresponds to a sweep occurring within the past ~4,000 years. Finally, we performed Bayesian model selection

to compare a neutral model ($s = 0$) with a model that included selection ($0.001 < s < 0.1$; Supplementary Methods [Supplementary Material online]). The posterior probability of the neutral model was 0.037, while that of the selection model was 0.963, indicating that the patterns of DNA sequence polymorphism observed in the *CG9509* region are much better explained by a combination of selection and demography than by demography alone.

Functional analysis of the *CG9509* upstream region

To test whether or not sequence variants in the *CG9509* upstream region have an effect on gene expression, we created two reporter gene constructs in which 1.2 kb of the *CG9509* upstream region from either a European or an African strain was fused to the *Escherichia coli lacZ* gene. The tested region begins just after the *CG14406* stop codon and ends just before the *CG9509* start codon (fig. 1). The reporter genes were inserted into a common genetic background using phiC31 site-specific integration (Groth et al. 2004; Bischof et al. 2007), which allowed us to target the reporter genes to specific sites in the genome and compare the expression driven by the *CG9509* upstream sequences in an otherwise identical genetic background.

In both males and females we found that the European version of the *CG9509* upstream region drives significantly higher expression than the African version, with the average expression difference being ~3.5-fold (fig. 4). The results were consistent for two independent phiC31 integration sites (68E and 86Fb) and whether reporter gene expression was measured at the level of protein (β-galactosidase activity) or transcript (qRT-PCR) abundance (fig. 4).

The above expression assays, like the original microarray experiments, used whole flies as the source material. Because the native *CG9509* gene is known to show highly enriched expression in Malphigian tubule (Chintapalli et al. 2007), we examined reporter gene expression specifically in this tissue. Staining of β-galactosidase activity confirmed that the reporter gene had enriched expression in Malphigian tubule (Supplementary fig. S1 [Supplementary Material online]) and enzyme activity assays revealed that the European version of the *CG9509* upstream region drives ~3-fold higher expression than the African version in Malphigian tubule (fig. 4). These results indicate that the increased expression observed for the European upstream region is not a result of it driving broader expression across multiple tissues, but of it driving quantitatively greater expression in the tissue where it is normally expressed.

Discussion

Several lines of evidence suggest that *CG9509* has been a target of adaptive *cis*-regulatory evolution in non-African *D. melanogaster*. First, this gene shows large and consistent differences in expression between Zimbabwe and cosmopolitan flies. Second, nucleotide polymorphism in the upstream region is greatly reduced in a European sample of alleles, which is consistent with a recent selective sweep. Third, a portion of the upstream region that co-localizes with the putative selective sweep is sufficient to recapitulate expression differences between European and African alleles in an *in vivo* reporter gene assay. Although our expression and sequence analyses cover only a single population from the Netherlands, other data suggest that the selective sweep of the *CG9509* upstream region is common to many non-African populations. At the expression

level, a large difference in *CG9509* expression was reported between non-African (North America and Asian) and African (Zimbabwe) strains (Meiklejohn et al. 2003; Table 1). At the DNA sequence level, two observations support a selective sweep outside of Africa in the *CG9509* upstream region. First, the monomorphic upstream region that we identified in our European sample is identical to the *D. melanogaster* reference sequence (Adams et al. 2000), which comes from a non-African lab strain. Second, the same region is also monomorphic (and identical in sequence to our Netherlands sample) in a pooled sample of 113 isofemale lines from Portugal (Pandey et al. 2011).

Selection appears to have favored a quantitative increase in *CG9509* expression, rather than a qualitative change in spatial pattern of expression. According to FlyAtlas (Chintapalli et al. 2007), *CG9509* shows highly enriched expression in Malphigian tubule, with a ratio of Malphigian tuble to whole body expression of 24. Since the FlyAtlas data were generated from a non-African fly strain, it is unlikely that the expression difference between populations could be explained by an increase in the breadth of *CG9509* expression. Consistent with this, we find that the difference in Malphigian tubule expression driven by the European and African versions of the *CG9509* upstream region is indistinguishable from the difference observed in whole flies (fig. 4).

The relative difference in *CG9509* expression between populations is the same in males and females. This may be unusual, as previous studies have shown that the vast majority of expression differences between the European and African populations are sex-specific (Hutter et al. 2008; Müller et al. 2011). However, when comparing the two sexes, the overall expression of *CG9509* is higher in males than females (Gnad and

Parsch 2006). We also see this pattern in our reporter gene assays, where the expression level (measured by β-galactosidase activity) in males is ~20% higher than in females. It is noteworthy that the insecticide resistance gene, *Cyp6g1*, also shows a large between-population expression difference in both sexes, enriched expression in Malphigian tubule, and an overall pattern of male-biased expression (Gnad and Parsch 2006). Although the exact function of *CG9509* is unknown, its expression profile and annotation as a choline dehydrogenase suggest that, like *Cyp6g1*, it may play a role in detoxification. Alternatively, because the ratio of phosphatidylcholine to phosphatidyethanolamine is known to decrease during cold acclimation (Kostal et al. 2011) or in the presence of dietary ethanol (Miller et al. 1993), it may be that *CG9509* plays a role in regulating phospholipid content and that the optimal ratio of these two phospholipids differs between African and non-African environments.

The polymorphism data allow us to narrow the putative target of selection to a 1.8-kb region that is monomorphic within Europe. The small size of the chromosomal region affected by the selective sweep is likely a result of there being a relatively high recombination rate (4.8×10^{-8} per bp per generation) in this part of the X chromosome (Comeron et al. 1999). The size of the sweep also depends on the strength of selection and the time since the sweep began. However, it is not possible to estimate the selection coefficient (s) and the time of the sweep (t) accurately (fig.3). Importantly, simulations with strong selection ($0.01 < s < 0.1$) and recent selective sweeps ($t < 30{,}000$ generations) indicate that the median value of segregating sites in the genomic region is 48 (with 5th and 95th quantiles of 3 and 101, respectively). Therefore, observing 34 segregating sites in a region of 5.6 kb (as in our data) is expected even under strong and recent selection.

Our reporter gene assays allow us to further narrow the target of selection to a 1.2-kb region. Within this region, there are only two fixed differences between all European and all African alleles: one is a 5-bp indel that occurs 820 bp upstream of the *CG9509* start codon and the other is an A/T single nucleotide polymorphism (SNP) that occurs 1,150 bp upstream of the start codon. In addition, there are six other SNPs where a derived variant is fixed within the European sample, but at low frequency (8–25%) in the African sample. All of the above differ between the European and African alleles used in our reporter gene assays. Thus, the combination of population genetic and functional analyses has allowed us to restrict the candidate sites for the target of selection to a tractable number that can be examined individually in future experiments.

Supplementary Material

Supplementary methods and supplementary figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgements

We thank H. Gebhart, C. Iannitti, and Y. Cämmerer for excellent technical assistance in the lab. We also thank R. Hudson for providing computer code. This work was carried out as part of the research unit "Natural selection in structured populations" (FOR 1078) funded by *Deutsche Forschungsgemeinschaft* grant PA 903/5. P.P. was supported by grant I/824234 from the Volkswagen Foundation.

References

Adams MD, Celniker SE, Holt RA, et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.

Bachtrog D. 2008. Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evol Biol*. 8:334.

Bischof J, Maeda RK, Hediger M, Karch F, Basler K. 2007. An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. *Proc Natl Acad Sci USA* 104:3312–3317.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in Arabidopsis. *Nature* 416:531–534.

Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* 101:577–580.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134:25–36.

Catania F, Kauer MO, Daborn PJ, Yen JL, Ffrench-Constant RH, Schlotterer C. 2004. World-wide survey of an *Accord* insertion and its association with DDT resistance in *Drosophila melanogaster*. *Mol Ecol*. 13:2491–2504.

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

Chung H, Bogwitz MR, McCart C, Andrianopoulos A, Ffrench-Constant RH, Batterham P, Daborn PJ. 2007. *Cis*-regulatory elements in the *Accord* retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* 175:1071–1077.

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.

Csilléry K, Blum MGB, Gaggiotti OE, François O. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol*. 25:410–418.

Daborn PJ, Yen JL, Bogwitz MR, et al. (13 co-authors). 2002. A single p450 allele associated with insecticide resistance in *Drosophila*. *Science* 297:2253–2256.

David JR, Capy P. 1988. Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet*. 4:106–111.

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet*. 37:544–548.

Ewing, G, Hermisson J. 2010. MSMS: A coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics* 26:2064–2065.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 415:1024–1026.

Gilad Y, Oshlack A, Rifkin SA. 2006. Natural selection on gene expression. *Trends Genet*. 22:456–461.

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.

Gnad F, Parsch J. 2006. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics* 22:2577–2579.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433:481–487.

Grath S, Baines JF, Parsch J. 2009. Molecular evolution of sex-biased genes in the *Drosophila ananassae* subgroup. *BMC Evol Biol*. 9:291.

Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of transgenic *Drosophila* by using the site-specific integrase from phage phiC31. *Genetics* 166:1775–1782.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res*. 15:790–799.

Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol Biol Evol*. 25:1825–1834.

Hense W, Baines JF, Parsch J. 2007. X chromosome inactivation during *Drosophila* spermatogenesis. *PLoS Biol*. 5:e273.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61:995–1016.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

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.

Hutter S, Saminadin-Peter SS, Stephan W, Parsch J. 2008. Gene expression variation in African and European populations of *Drosophila melanogaster*. *Genome Biol*. 9:R12.

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. 2008. The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. *Cell* 132:783–793.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 60:765–777.

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

King MC, Wilson AC.1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.

Kostal V, Korbelova J, Rozsypal J, Zahradnickova H, Cimlova J, Tomcala A, Simek P. 2011. Long-term cold acclimation extends survival time at 0°C and modifies the metabolomic profiles of the larvae of the fruit fly *Drosophila melanogaster*. PLoS One 6:e25025.

Lemos B, Meiklejohn CD, Cáceres M, Hartl DL. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* 59:126–137.

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

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Maside X, Charlesworth B. 2007. Patterns of molecular variation and evolution in *Drosophila americana* and its relatives. *Genetics* 176:2293–2305.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL. 2007. Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature* 448:587–590.

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc Natl Acad Sci USA* 100:9894–9899.

Miller RR Jr, Yates JW, Geer BW. 1993. Dietary ethanol reduces phosphatidylcholine levels and inhibits the uptake of dietary choline in *Drosophila melanogaster* larvae. Comp Biochem Physiol Comp Physiol.104:837–844.

Müller L, Hutter S, Stamboliyska R, Saminadin-Peter SS, Stephan W, Parsch J. 2011. Population transcriptomics of *Drosophila melanogaster* females. *BMC Genomics* 12:81.

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.

Pandey RV, Kofler R, Orozco-terWengel P, Nolte V, Schlötterer C. 2011. PoPoolation DB: a user-friendly web-based database for the retrieval of natural polymorphisms in *Drosophila*. *BMC Genet*. 12:27.

Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in *Drosophila*. *Mol Biol Evol*. 26:691–698.

Pavlidis P, Jensen JD, and Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from non-equilibrium populations. *Genetics* 185:907–922.

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB. 2006. Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature* 440:1050–1053.

Rifkin SA, Houle D, Kim J, White KP. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* 438:220–223.

Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc Natl Acad Sci USA* 104:6504–6510.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.

Stephan W, Li H. 2006. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98:65–68.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* 172:1607–1619.

Townsend JP, Hartl DL. 2002. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol*. 3:RESEARCH0071.

Tweedie S, Ashburner M, Falls K, et al. (12 co-authors). 2009. FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res*. 37:D555–D559.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet*. 8:206–216.

Table 1. Expression Divergence of *CG9509* Between African and Non-African Strains

Method	Comparison	Sex	Ratio	Source
Microarray	Cosmopolitan *vs*. Zimbabwe	Male	2.7***	Meiklejohn et al. (2003)
Microarray	Europe *vs*. Zimbabwe	Male	2.3***	Hutter et al. (2008)
qRT-PCR	Europe *vs*. Zimbabwe	Male	2.0**	Present study
qRT-PCR	Europe *vs*. Zimbabwe	Female	3.2**	Present study

The "Ratio" column gives the average ratio of non-African-to-African expression. Statistical significance of microarray data was assessed by a Bayesian method (Townsend and Hartl 2002), while that of qRT-PCR data was determined by a *t*-test using mean expression values from eight strains of each population (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

Table 2. Results of McDonald-Kreitman (MK) Tests for *CG9509*

							MK-test *P*-value	
Population	D_s	P_s	D_n	P_n	D_{up}	P_{up}	**Nonsynonymous**	**Upstream**
Europe	44	7	51	3	84	1	0.1935	0.0045
Africa	40	40	49	8	76	33	0.0001	0.0067

Shown are the numbers of fixed differences (*D*) between each *D. melanogaster* population and *D. sechellia* and the number of polymorphic sites (*P*) within *D. melanogaster*. Subscripts indicate synonymous (s), nonsynonymous (n), and upstream (up) sites. The upstream region spans the entire sequence between the *CG14406* and *CG9509* coding regions (see fig. 1).

Figure legends

Fig. 1. Polymorphism and divergence in the *CG9509* genomic region. Nucleotide diversity (θ) in the European (black line) and African (dotted line) populations, as well as the mean pairwise divergence between all *D. melanogaster* sequences and *D. sechellia*, are shown for a sliding-window analysis (window size = 200 bp, step size = 25 bp). The positions of the three transcriptional units contained within this region are shown below, with solid boxes representing exons and open boxes representing introns. The arrowhead indicates the direction of transcription. The hatched box indicates the portion of the *CG9509* upstream region used for reporter gene analysis.

Fig. 2. Classification of the *CG9509* region as either neutral or selected. (A) Machine-learning classification surface on the basis of two test statistics *SweepFinder* (SF; Nielsen et al. 2005) and ω (Kim and Nielsen 2004). A support vector machine (SVM) algorithm implemented in the 'e1071' R package and 16,000 training points (not shown) were used for the construction of the classification surface. Training points were generated by simulation. The gray region denotes neutrality, whereas the white region denotes selection. The observed data point is indicated. The false positive rate is 28% for the entire feature space (see also text). Only the feature space associated with 85% of the simulated points is shown, which are more relevant to the observed data point (x-axis: 0–70, y-axis: 0–27). Large values of SF and ω are not shown, but were considered in the entire training set. (B) The probability that k or more selection instances are obtained among x simulations for a given number of points (distance) around the observed point. For the limited area around the observed data point, this probability is very low ($\sim 10^{-8}$),

indicating a significant over-representation of selection instances in this region of the feature space.

Fig. 3. Bayesian estimation of selective sweep parameters. Dotted lines indicate the prior distributions, gray lines indicate the distributions obtained after the rejection step, and solid black lines indicate the posterior distributions (rejection plus local linear regression). (A) The posterior distribution of the selection coefficient, s. The prior probability of neutrality ($s = 0$) is 0.5, whereas the posterior probability of the neutral model is 0.037. Although there is strong evidence of $s > 0$, it is not possible to infer the value of the selection coefficient accurately. (B) The posterior distribution of the time of the sweep, t. The prior distribution of t is uniform from the split of the European and African populations to the present. The posterior distribution suggests a relatively recent sweep occurring within the past 100,000 generations (~10,000 years).

Fig. 4. Ratio of reporter gene expression driven by the European *CG9509* upstream region to that driven by the African upstream region. Assays were performed separately on whole males, whole females, and dissected (male) Malphigian tubules for two independent reporter gene insertion sites (68E and 86Fb). For males and females, expression was measured at the level of both protein (β-galactosidase activity) and mRNA (qRT-PCR) abundance. Error bars indicate the 95% confidence interval. In all cases, the European upstream region drove significantly higher expression than the African upstream region (t-test, $P < 0.025$).

Fig. 1

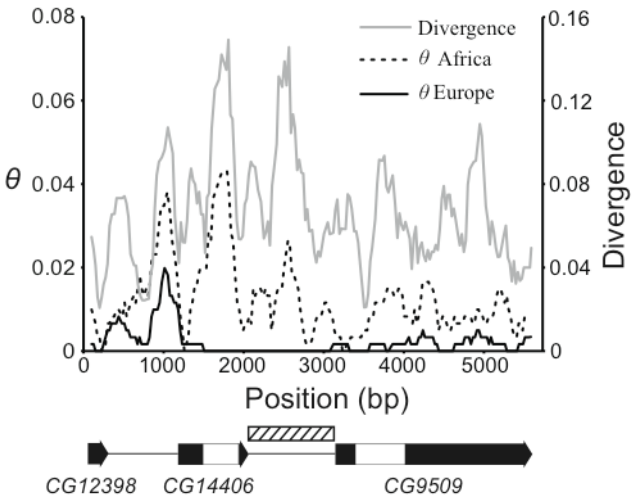

Fig. 2

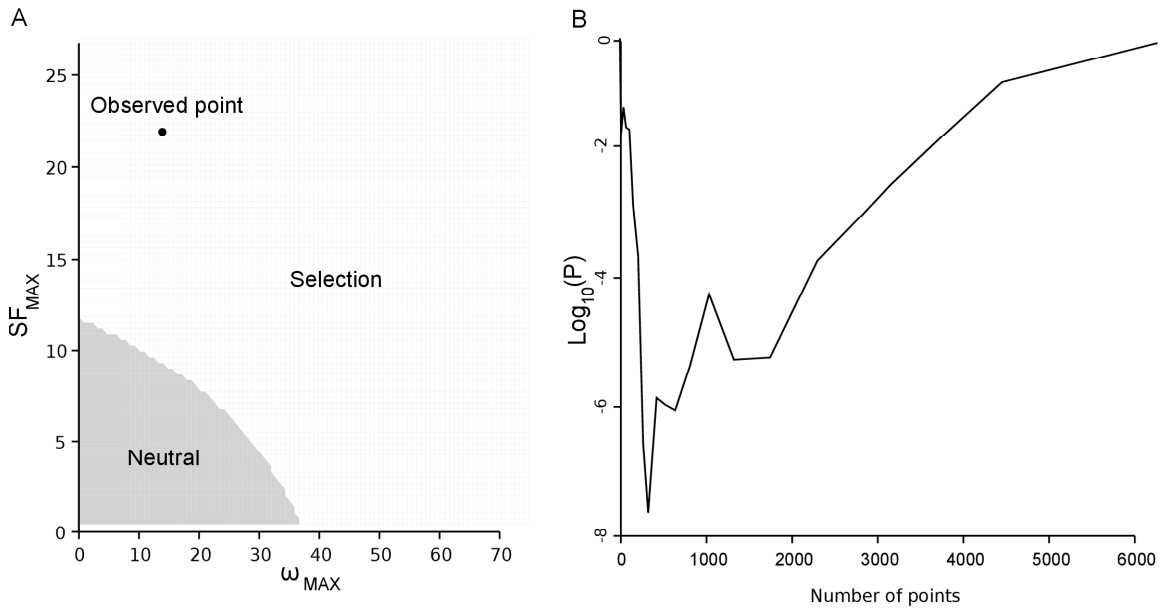

Fig. 3

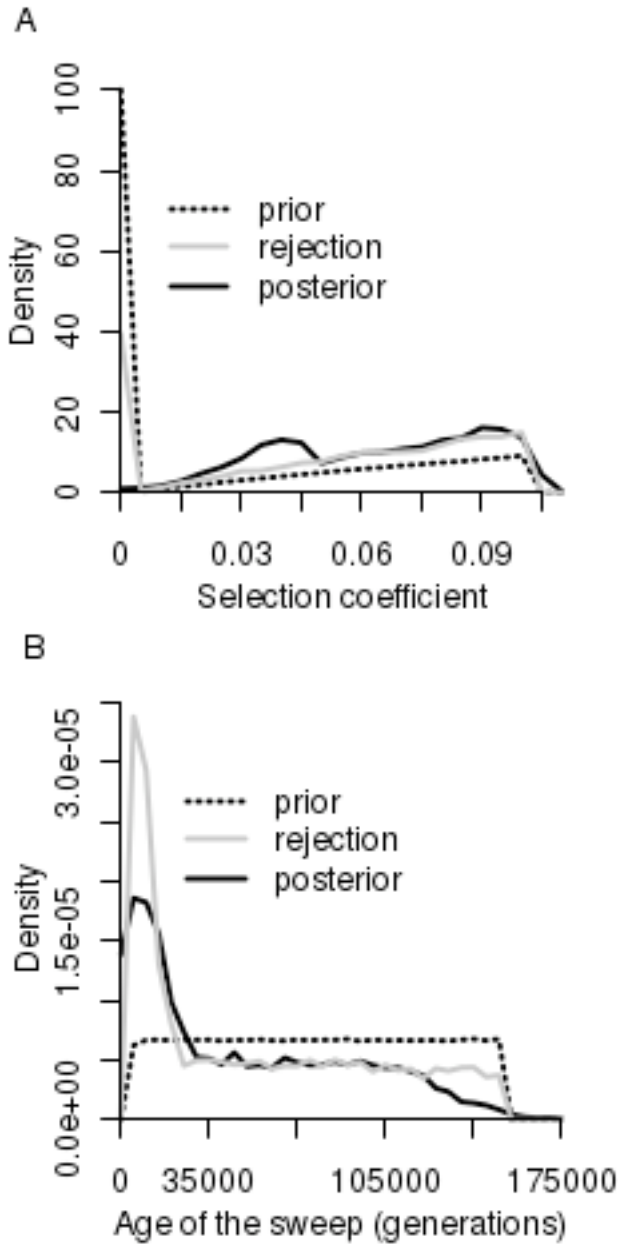

Fig. 4