Adaptive divergence of a transcriptional enhancer between populations of Drosophila melanogaster

Amanda Glaser-Schmitt†, Ana Catalán† and John Parsch

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

As species colonize new habitats they must adapt to the local environment. Much of this adaptation is thought to occur at the regulatory level; however, the relationships among genetic polymorphism, expression variation and adaptation are poorly understood. Drosophila melanogaster, which expanded from an ancestral range in sub-Saharan Africa around 15 000 years ago, represents an excellent model system for studying regulatory evolution. Here, we focus on the gene CG9509, which differs in expression between an African and a European population of D. melanogaster. The expression difference is caused by variation within a transcriptional enhancer adjacent to the CG9509 coding sequence. Patterns of sequence variation indicate that this enhancer was the target of recent positive selection, suggesting that the expression difference is adaptive. Analysis of the CG9509 enhancer in new population samples from Europe, Asia, northern Africa and sub-Saharan Africa revealed that sequence polymorphism is greatly reduced outside the ancestral range. A derived haplotype absent in sub-Saharan Africa is at high frequency in all other populations. These observations are consistent with a selective sweep accompanying the range expansion of the species. The new data help identify the sequence changes responsible for the difference in enhancer activity.

1. Introduction

(a) The importance of gene regulation in adaptation

Differences in gene expression are thought to underlie many of the phenotypic differences between species and populations [1–3]. With the advent of transcriptomic technologies, such as microarrays and high-throughput RNA sequencing (RNA-seq), it has become possible to identify the genes that differ in expression between species or vary in expression among individuals of the same species. Such studies have revealed that there is considerable expression divergence between closely related species (e.g. human and chimpanzee [4] or Drosophila melanogaster and Drosophila simulans [5]) as well as abundant expression variation within species (e.g. human [4,6,7], mouse [8], Drosophila [9,10], yeast [11–13] and fish [14–16]). A current challenge in evolutionary genetics is to identify the specific genetic changes responsible for differences in gene expression and to determine how these changes impact an organism’s fitness. In this context, much attention has been paid to cis-regulatory elements, such as transcriptional enhancers, as they are known to play a key role in regulatory evolution [17]. It has been argued that cis-regulatory evolution is the major driver of adaptive divergence between species, especially at the level of morphology [17–19]. However, the importance of cis-regulatory divergence in relation to other types of genetic changes (e.g. amino acid replacements within proteins) in adaptation is still a topic of debate [20].

A well-known example of adaptive cis-regulatory evolution in humans involves the lactase gene (LCT), where single-nucleotide polymorphisms (SNPs) in an upstream regulatory element are associated with persistent expression of LCT in adults and enable them to digest the milk sugar lactose [21]. Patterns
of DNA sequence polymorphism in the LCT region suggest that it has been the target of recent positive selection within northern European populations [22]. Furthermore, the discovery of different, independently derived SNPs in this region of the genome that are associated with lactase persistence in African pastoralist populations is indicative of convergent adaptive evolution [23]. In D. melanogaster, polymorphism in the expression of the cytochrome P450 gene Cyp6g1 is associated with the insertion of an Accord transposable element into its upstream regulatory region [24]. Overexpression of Cyp6g1 owing to the Accord insertion confers resistance to the insecticide DDT [25], a trait that is in high frequency in non-African populations [26]. Patterns of DNA sequence polymorphism are consistent with recent positive selection favouring the high-expression allele [26]. The Cyp6g1 example illustrates how the powerful genetic resources available for D. melanogaster can be used to identify adaptive changes in gene expression.

(b) The demographic history of Drosophila melanogaster

Drosophila melanogaster is currently a cosmopolitan species with a worldwide distribution [27]. However, the global spread of the species from its ancestral range in sub-Saharan Africa is thought to have occurred relatively recently [27,28]. Genomewide analyses of DNA sequence variation in multiple African and non-African populations have resulted in our current understanding of the species’ biogeographic and demographic history [29–33]. A general pattern that has been observed is that DNA sequence polymorphism is greater among individuals from sub-Saharan Africa than among individuals from other worldwide locations [29,34–36], which is consistent with an Afrotropical origin of the species. Populations from southern-central Africa (e.g. Zambia and Zimbabwe) show the highest genetic diversity, suggesting that they best represent the centre of origin [32]. It is hypothesized that the initial expansion of D. melanogaster from its ancestral range occurred around 15 000 years ago with the colonization of human settlements in the Middle East [31]. The colonization of Europe and Asia from this original non-African source population is thought to have occurred more recently, within the past 2500–5000 years and been concomitant with the spread of human populations and agriculture [31]. Finally, the colonization of North America is documented to have occurred within the past 200 years [37] and appears to have involved the admixture of European and African D. melanogaster [33]. There is also evidence for recent non-African gene flow into sub-Saharan Africa, with the extent of admixture varying among African populations [32].

(c) Expression differences between Drosophila melanogaster populations

Its successful colonization of non-African territories suggests that D. melanogaster has undergone adaptation to new environmental conditions. Given our extensive knowledge of the D. melanogaster genome and its tractability as a model organism, there has been considerable interest in finding the genes and genetic changes that underlie this adaptation. One approach has been to look for regions of the genome that show patterns of sequence polymorphism indicative of recent positive selection [38,39]. These studies have identified genes or regions of the genome that are candidates for adaptive evolution [29,30,32,40], but in most cases it has been difficult to link genetic variants with functional or phenotypic differences between populations. Another approach has been to look for genes that differ in expression between African and non-African flies. This approach focuses on regulatory divergence.

To date, such expression studies have been carried out using whole adult males [9,41], whole adult females [42] and the dissected brains of both sexes [43]. In all of these cases, hundreds of genes differing in expression between populations were identified. However, the overlap among the differentially expressed genes identified by each study was small, suggesting that regulatory evolution often occurs in a sex- and tissue-dependent fashion [42,43].

(d) Population genetics and expression of CG9509

One gene that shows a large and consistent expression difference between African and non-African flies of both sexes is CG9509 [9,41,44]. The specific function of this gene in D. melanogaster is unknown, although sequence homology has led to it being annotated as a choline dehydrogenase [45]. In addition, its highly enriched expression in the Malpighian tubules [46] suggests that it may play a metabolic role in detoxification. The sequence and expression of CG9509 have been studied in detail in population samples from Europe (The Netherlands) and Africa (Zimbabwe), revealing three major features [44]. First, CG9509 shows two to three times higher expression in the European population than in the African population (figure 1). Second, sequence polymorphism in the CG9509 region is greatly reduced in the European population, especially in the intergenic region just upstream of the CG9509 coding sequence, which is consistent with a recent selective sweep. Third, this intergenic region (here denoted as the CG9509 enhancer) is sufficient to drive differences in reporter gene expression equal to those observed for the CG9509 gene in natural populations (figure 1). Taken together, these results provide strong evidence that positive selection has acted on the CG9509 enhancer to increase expression in the European population. To better understand the timing and geographical scale of this positive selection, we extended the analysis of the CG9509 enhancer to new population samples from Europe, Asia, northern Africa and sub-Saharan Africa. We find that sequence polymorphism is very low in all
populations outside the ancestral range, but much higher within sub-Saharan Africa. Furthermore, a derived haplotype associated with elevated CG9509 expression is at high frequency in all populations outside sub-Saharan Africa but was not detected within the ancestral range. These results suggest that selection for increased expression of CG9509 occurred during or soon after the out-of-Africa expansion of the species, before its spread into Europe and Asia.

2. Material and methods

(a) Population samples

Sequence polymorphism was surveyed in the following six *D. melanogaster* populations samples: 12 isofemale lines from The Netherlands (Leiden), 11 isofemale lines from Germany (Munich), 11 isofemale lines from Malaysia (Kuala Lumpur), 12 isofemale lines from Egypt (Cairo), 10 isofemale lines from Zambia (Siavonga) and 12 isofemale lines from Zimbabwe (Lake Kariba). The Zimbabwe and The Netherlands populations were used in a previous study of sequence and expression variation associated with the CG9509 enhancer region [44], as well as in previous genome-wide studies [29,35,36,47]. The Malaysian population also was used in previous genome-wide demographic studies [31,48]. At least six strains from each population were used for quantitative reverse-transcription PCR (qRT-PCR) analysis. Flies from all populations were maintained as inbred, isofemale lines under standard conditions (22°C, 14 L : 10 D cycle, cornmeal-molasses medium) for at least 10 generations prior to expression analyses.

(b) DNA sequencing

New sequences of the CG9509 intergenic region were obtained from isofemale lines of the German, Malaysian, Egyptian and Zambian populations. For each line, DNA was extracted from a single male fly using the MasterPure DNA Purification Kit (Epicentre). PCR was performed under standard conditions using four primer pairs published in Samadinin-Peter et al. [44] and one additional reverse primer (5′-AGCTCAGACAGCAGACCGTAT-3′). The amplified region consisted of 1.2 kb of intergenic sequence, ranging from the stop codon of CG14406 to the start codon of CG9509. PCR products were purified with ExoSAP-IT (USB) and sequenced using BigDye chemistry on a 3730 automated sequencer (Applied Biosystems). Both strands of DNA were sequenced using the PCR primers as sequencing primers. Trace files were edited using SEQTRACE [49] and a multiple sequence alignment was generated with SEAVIEW (v. 4 [50] using the ClustalW2 algorithm. All sequences have been submitted to the GenBank/EMBL database under the accession numbers HP913659–HP913726.

(c) Population genetic analyses

The following summary statistics were calculated using DNAsp v. 5.10.1 [51]: mean pairwise nucleotide diversity (π), Watterson’s estimate of nucleotide diversity (θ) [52], number of segregating sites, haplotype number, haplotype diversity, Fst and Dxy (average pairwise differences between populations). Within each population, the 95% CIs of π and θ were estimated from 10 000 coalescent simulations. A neighbour-joining tree of all sequences was constructed using MEGA v. 5.05 [53]. For this, the evolutionary distances were calculated using the maximum composite likelihood method. Clade support was assessed from 1000 bootstrap replicates.

To determine whether the observed features (number of segregating sites, number of haplotypes and number of fixed, derived variants) in the populations outside sub-Saharan Africa could be explained solely by an out-of-Africa bottleneck, we performed coalescent simulations with ms [54], using bottleneck parameters inferred previously for the X chromosome [31,55]. To match the structure of our observed data, we simulated samples from two present-day populations of sizes N and 0.34N, with sample sizes of 22 and 46 sequences, respectively. The larger sample was drawn from a population that experienced a bottleneck approximately 15 000 years ago, which reduced the population to 0.5% of its ancestral size. The smaller sample was drawn from a population that maintained a constant population size. Prior to the bottleneck, the two populations were assumed to be part of a single panmictic population of size N. Simulations were conditioned on the observed number of segregating sites in the total sample with a local recombination rate of 3.47 cm/Mb [56]. A total of 100 000 simulations were performed and the p-value was determined as the proportion of simulated datasets in which one of the above features in the bottlenecked population (46 sequences) was equal to (or more extreme than) the observed value in the combined non-sub-Saharan African populations.

(d) Expression analysis

Total RNA was extracted from 10 to 15 adult males (aged 4–6 days) and DNase I digestion was performed using the MasterPure RNA Purification Kit (Epicentre). For each strain, at least two biological replicates were performed. For each replicate, 3 μg total RNA was reverse-transcribed using random hexamer primers and Superscript II reverse transcriptase (Invitrogen) following the manufacturer’s protocol. A TaqMan Gene Expression Assay (Invitrogen) was then performed on the resulting cDNA using a probe specific to CG9509 (Dm01838873_g1) as well as a probe specific to the ribosomal protein gene RpL32 (Dm02151827_g1), which was used as an endogenous control. Since the amplification efficiencies of the two probes were nearly identical (within the range 96–99%), the ΔΔCt method was used to calculate normalized gene expression [57]. Briefly, the average threshold cycle (Ct) was determined for two technical replicates per biological replicate and ΔCt was calculated as the mean Ct difference between the CG9509 and RpL32 probes. The fold-change difference in expression for each biological replicate relative to the Zimbabwe population was then calculated as 2^-[ΔΔCt], where ΔCt is the mean ΔCt value for each biological replicate and ΔΔCt is the mean ΔCt value of the Zimbabwe strains. In order to ensure a balanced design, a total of six strains per population, each with two biological replicates, was used. For strains where more than two biological replicates were performed, the two replicates with ΔCt closest to the median were used.

3. Results

(a) Sequence polymorphism in the CG9509 enhancer

A previous population genetic analysis of the CG9509 enhancer examined only one population from Europe (The Netherlands) and one population from sub-Saharan Africa (Zimbabwe) [44]. To obtain a broader view of genetic variation, we sequenced the 1.2 kb intergenic region between CG9509 and CG14406 (figure 2) in new populations samples from Europe (Germany), Asia (Malaysia), northern Africa (Egypt) and sub-Saharan Africa (Zambia). In the following, we refer to the populations from outside sub-Saharan Africa as ‘cosmopolitan’. Overall, we find that nucleotide diversity is very low in all the cosmopolitan populations (mean θ of 0.07%), with many individuals sharing the same haplotype (table 1). By contrast, nucleotide diversity is at least 12-fold higher in the Zambia and Zimbabwe
or the expression difference between European and African alleles.

The intergenic region between the stop codon of CG14406 and the start codon of CG9509 was used for the population genetic analysis. This region has been shown to contain the transcriptional enhancer responsible for the expression difference between European and African alleles.

To determine whether the reduction in polymorphism observed in the cosmopolitan populations could be explained solely by an out-of-Africa bottleneck, we performed coalescent simulations using a demographic model inferred from X chromosome-wide polymorphism data [31,55]. Of 100 000 simulated datasets, none showed a reduction in θ as great as that observed in the real data, indicating that the probability of it being caused by a bottleneck alone is less than 0.00001. Two other features of the observed data, the number of haplotypes and the number of derived variants fixed in the cosmopolitan populations, were also highly unlikely to have been caused by a bottleneck alone (p < 0.00001).

(b) Sequence divergence between populations

For the cosmopolitan populations, there is not only low sequence diversity within each population, but also very little sequence divergence between populations. On average, $F_{at}$ is 0.09 among these populations, while the average pairwise nucleotide divergence between populations ($D_{spa}$) is 0.08% (see electronic supplementary material, table S1). By contrast, these populations show much greater sequence divergence than the sub-Saharan African populations, with $F_{at}$ averaging 0.46 and $D_{spa}$ averaging 1.12%. There is little sign of population structure between the Zambia and Zimbabwean populations, where $F_{at}$ is 0.001. The above features are also evident in a neighbour-joining tree, where the cosmopolitan lineages form an exclusive clade with very short branch lengths (figure 3), suggesting that they descend from a very recent common ancestor. By contrast, the Zambian and Zimbabwean sequences are separated by longer branches, which is consistent with an older age of these alleles (figure 3).

(c) Sequence differences between cosmopolitan and sub-Saharan African populations

Experiments using a transgenic reporter gene have shown that the twofold to threefold CG9509 expression difference observed between flies from The Netherlands and Zimbabwe is caused by sequence variation in a 1.2-kb enhancer located just upstream of the CG9509 coding region (figure 1) [44]. Within this region, there are nine sites that show a fixed or nearly fixed difference between the cosmopolitan and the sub-Saharan African populations (figure 4). These include eight SNPs and one insertion/deletion (indel) polymorphism. Using D. simulans, Drosophila sechellia and Drosophila yakuba as outgroup species, the ancestral state could be inferred for all eight SNPs (figure 4). In all cases, the sub-Saharan African variant was the ancestral form, indicating that new mutations have risen to high frequency in the other populations. For the indel polymorphism, it was not possible to determine the ancestral state, as multiple, large indels have occurred across this region in the outgroup species. However, the tight linkage of this indel polymorphism with the surrounding SNPs suggests that it represents a deletion mutation and that a common derived haplotype is present in all cosmopolitan populations. One strain from Zambia has a deletion similar to the one observed outside sub-Saharan Africa (figure 4). However, this may represent an independent mutational event, as there is also a unique SNP directly adjacent to the deletion in this strain (figure 4). Consistent with this interpretation, the deletion in the Zambia strain is not linked to any of the derived SNPs found at high frequency in the cosmopolitan populations (figure 4).

(d) Expression differences between populations

It was shown previously that CG9509 has higher expression in a cosmopolitan population (The Netherlands) than in a sub-Saharan African population (Zimbabwe; figure 1) [41,44]. Using qRT-PCR, we were able to confirm this result and extend it to three new cosmopolitan populations (Germany, Malaysia and Egypt) and a new sub-Saharan African population (Zambia). On average, the cosmopolitan strains showed nearly threefold higher expression than the sub-Saharan African strains, which was highly significant (figure 5). We also compared CG9509 expression in each cosmopolitan population to that in sub-Saharan Africa. Since the Zambian and Zimbabwean populations showed no evidence of population structure (see electronic supplementary material, table S1) and had very similar CG9509 expression (figure 5), they were pooled for comparison with the cosmopolitan populations. Individually, the populations from The Netherlands, Malaysia and Egypt each had significantly higher CG9509 expression than the pooled sub-Saharan African populations (figure 5). The German population showed, on average, 1.6-fold higher CG9509 expression than the pooled sub-Saharan African populations, but this difference was not significant (figure 5).

(e) Association between sequence variants and expression

To determine whether particular sites within the CG9509 enhancer that show a fixed or nearly fixed difference between cosmopolitan and sub-Saharan African populations (figure 4) were associated with the observed difference in expression, we examined the expression of CG9509 in additional strains from Zambia. However, we could not establish a clear link between any individual sequence variant and the expression difference. For example, Zambia strain Z173, which is the only sub-Saharan African strain with the 5-bp deletion at positions 821–817 before the CG9509 start codon (figure 4), did not show higher expression than the other sub-Saharan strains (see electronic supplementary material, figure S1). Similarly, strain Z112, which has cosmopolitan variants at positions 1180, 1174 and 1155, and strain Z1254, which has cosmopolitan variants at positions 748 and 718 (figure 4), did
Table 1. DNA sequence polymorphism within populations. n, number of sequences; S, number of segregating sites; θ, Watterson’s [52] estimate of nucleotide diversity (per 100 sites); π, mean pairwise nucleotide diversity (per 100 sites); nHap, number of haplotypes; HapDiv, haplotype diversity. The 95% CIs for θ and π are shown in brackets.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>S</th>
<th>θ</th>
<th>π</th>
<th>nHap</th>
<th>HapDiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Netherlands</td>
<td>12</td>
<td>1</td>
<td>0.03 (0.00 – 0.13)</td>
<td>0.05 (0.00 – 0.16)</td>
<td>2</td>
<td>0.55</td>
</tr>
<tr>
<td>Germany</td>
<td>11</td>
<td>5</td>
<td>0.14 (0.00 – 0.27)</td>
<td>0.11 (0.00 – 0.31)</td>
<td>5</td>
<td>0.71</td>
</tr>
<tr>
<td>Malaysia</td>
<td>11</td>
<td>1</td>
<td>0.03 (0.00 – 0.14)</td>
<td>0.05 (0.00 – 0.15)</td>
<td>2</td>
<td>0.55</td>
</tr>
<tr>
<td>Egypt</td>
<td>12</td>
<td>3</td>
<td>0.08 (0.00 – 0.24)</td>
<td>0.09 (0.00 – 0.27)</td>
<td>4</td>
<td>0.76</td>
</tr>
<tr>
<td>Zambia</td>
<td>10</td>
<td>42</td>
<td>1.33 (0.31 – 2.06)</td>
<td>1.03 (0.27 – 2.32)</td>
<td>10</td>
<td>1.00</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>12</td>
<td>39</td>
<td>1.12 (0.29 – 1.67)</td>
<td>0.87 (0.23 – 1.95)</td>
<td>12</td>
<td>1.00</td>
</tr>
</tbody>
</table>

not show unusually high expression relative to other Zambian strains (see electronic supplementary material, figure S1).

Although the German population showed lower average CG9509 expression than the other cosmopolitan populations (figure 5), this difference was not caused solely by strains MU10 and MU11, which were the only ones with the sub-Saharan variant (G) at position 167 (figure 4 and electronic supplementary material, figure S1). Within the cosmopolitan populations, there is a SNP (a G/C polymorphism 67 bp before the CG9509 start codon) segregating at intermediate frequency (32%; see electronic supplementary material, figure S2). The derived variant (G) is associated with a 1.5-fold increase in CG9509 expression within cosmopolitan populations (t-test; p = 0.016; see electronic supplementary material, figure S3). While this variant can account for some of the CG9509 expression variation among cosmopolitan strains, it cannot account for the large expression difference between cosmopolitan and sub-Saharan African strains, as cosmopolitan strains with the sub-Saharan African variant (C) still have over twofold higher expression than sub-Saharan African strains (t-test; p < 10^{-3}; see electronic supplementary material, figure S3).

4. Discussion

(a) Evidence for adaptive evolution of CG9509 at the level of expression

Several lines of evidence suggest that CG9509 has undergone adaptive regulatory evolution within the past 5000–15,000 years. First, this gene shows a large and consistent expression difference between cosmopolitan and sub-Saharan African populations (figure 5) [9,41,44]. Second, within cosmopolitan populations, DNA sequence polymorphism is greatly reduced in the intergenic region immediately upstream of the CG9509 coding sequence (table 1), which is consistent with a selective sweep in this region of the genome [44]. Third, sequence variation within this intergenic region (designated as the CG9509 enhancer) has been shown to account for the difference in expression between cosmopolitan and sub-Saharan African strains [44]. Finally, within the CG9509 enhancer, there is a derived haplotype that is in high frequency in cosmopolitan populations, but is absent in sub-Saharan Africa (figure 4).

The CG9509 enhancer also shows evidence for long-term adaptive evolution over the past 2–3 Myr (since the divergence of D. melanogaster and species of the D. simulans clade). Application of the McDonald-Kreitman (MK) test [58] to data on polymorphism within D. melanogaster and divergence between D. melanogaster and D. sechellia found a significant excess of between-species divergence in the enhancer compared to synonymous sites in the CG9509 coding region [44]. Although the previous analysis did not polarize divergence to the D. melanogaster lineage, a re-analysis of the data using D. yakuba as an outgroup to polarize changes indicated that a significant excess of substitutions in the enhancer occurred on the D. melanogaster lineage (see electronic supplementary material, table S2). This suggests that there have been recurrent selective sweeps within the D. melanogaster CG9509 enhancer since its divergence from D. sechellia.

(b) Evidence for adaptive evolution of CG9509 at the level of protein sequence

In addition to showing evidence for adaptive regulatory evolution, CG9509 also shows evidence for having undergone adaptive protein evolution within the past 2–3 Myr. A comparison of polymorphism and divergence within the CG9509 coding region using the MK test revealed a significant excess of non-synonymous divergence between species [44], which is indicative of recurrent selection for amino acid replacements. A recent genome-wide study of polymorphism also identified CG9509 as a target of positive selection using MK tests polarized to the D. melanogaster lineage [59]. Indeed, CG9509 was ranked among the top 10 genes in the genome that showed evidence for adaptive protein evolution on the D. melanogaster lineage [59].

(c) CG9509 sequence and expression variation within North America

Drosophila melanogaster is believed to have colonized North America within the past 200 years [37]. This colonization appears to be the result of admixture between European and African source populations, with the estimated proportion of European and African ancestry being 85% and 15%, respectively [33]. The Drosophila Genetic Reference Panel (DGRP) [60], consisting of 192 inbred, isofemale lines derived from a single outbred population from Raleigh, North Carolina, is an excellent resource for examining naturally occurring variation within a North American D. melanogaster population. Consistent with the inferred proportion of admixture in North America [33], the cosmopolitan variants at the sites showing fixed or nearly fixed differences between cosmopolitan and sub-Saharan African populations in the CG9509
enhancer (figure 4) are present in approximately 75–85% of the DGRP lines [60], while the private cosmopolitan variant (G 67 bp before the start codon; see electronic supplementary material, figure S2) is present in 31%.

The results of an association study of sequence and expression variation in a subset of 39 DGRP lines [61] are consistent with some of the major features of CG9509 sequence and expression variation identified in our study. First, in some DGRP lines the CG9509 enhancer region shows greatly reduced variant density in comparison to the surrounding regions [61], which is similar to the greatly reduced sequence polymorphism observed in our cosmopolitan strains (table 1). Second, DGRP lines showing this low variant density correspond to cosmopolitan haplotypes of the CG9509 enhancer that are associated with increased expression [44,61]. Third, the presence of cosmopolitan variants within the CG9509 enhancer region in particular DGRP lines appears to be associated with a general increase of CG9509 expression in these lines [61]. Analysis of the DGRP lines revealed an expression quantitative trait locus (eQTL) associated with CG9509 expression within the CG9509 enhancer region [61]. This eQTL corresponds to the segregating site 67 bp before the start codon (see electronic supplementary material, figure S2) that we found to be associated with CG9509 expression variation within cosmopolitan populations (see electronic supplementary material, figure S3). The direction and magnitude of the expression change [61] agree well with our finding that
the G variant at this site is associated with a 1.5-fold increase in expression within cosmopolitan populations (see electronic supplementary material, figure S3). However, none of the fixed or nearly fixed differences between cosmopolitan and sub-Saharan African populations (figure 4) showed a significant association with CG9509 expression within the DGRP lines [61]. This may be due to the fact that the analysis was performed on a single North American population in which sub-Saharan African variants were present only at low frequency, which reduces the statistical power to detect associations in genome-wide analyses.

(d) Possible functions of CG9509
At present, the specific function of CG9509 in D. melanogaster and the effect that variation in its expression has on phenotypic differences between individuals are unknown. CG9509 is predicted to encode a choline dehydrogenase with highly enriched expression in the Malpighian tubules [45,46], which is functionally analogous to the kidney of mammals. This suggests that CG9509 may play a role in detoxification. Variation in other genes involved in choline metabolism, namely choline kinases, has been implicated in insecticide resistance, with resistant alleles being present at high frequency in cosmopolitan D. melanogaster populations [43,62]. Unlike CG9509, these choline kinases show reduced expression (or loss of function) outside sub-Saharan Africa [43,59]. By contrast, resistance to DDT is conferred by overexpression of the cytochrome P450 gene Cyp6g1 [24], which also shows highest expression in the Malpighian tubules [46]. CG9509’s similarity in function and expression to these other insecticide resistance genes, as well as the strong signal for adaptive evolution outside sub-Saharan Africa, suggest that it may also play a role in the detoxification of insecticides or other chemicals present outside D. melanogaster’s ancestral home range.

It is also possible that CG9509 plays a role in adaptation to temperature or humidity. For example, it has been shown in Drosophila that the ratio of phosphatidylcholine to phosphatidylethanolamine decreases during cold acclimation [63], suggesting that choline metabolism might be linked to cold tolerance. Additionally, choline dehydrogenases are known to catalyse the conversion of choline into betaine [64], which has been reported to play an osmoprotectant role in mammals [65] and has also been found in insects [66]. CG9509’s very high expression in the Malpighian tubules (and lower expression in the gut) is consistent with a role in osmoregulation, which is a critical process for environmental adaptation. A QTL study of D. melanogaster did not find CG9509 to be among the major QTLs affecting desiccation resistance [67]. However, this study was carried out using recombinant inbred lines derived from two isofemale lines of a single North American (California) population and, thus, did not include genetic variation from sub-Saharan Africa.

Finally, knockout of the choline dehydrogenase gene (Chdh) in mice has been shown to decrease sperm motility [68]. Similarly, polymorphism in the human Chdh gene also is associated with variation in sperm motility [69]. Furthermore, dietary choline is required for proper sperm motility and reproductive behaviour in Drosophila [70]. Thus, it is possible that expression variation in the Drosophila CG9509 gene affects male fertility and/or sperm competition. Genes expressed in the testes, especially those that are X-linked, tend to show the greatest signal of adaptive evolution in Drosophila [71]. However, CG9509 shows only very low levels of expression in the testes that are several hundred-fold lower than those in the Malpighian tubules [46], making a role in male fertility unlikely.

5. Conclusion
Our finding that the selective sweep encompassing the CG9509 enhancer extends to populations from Asia and northern Africa has three important implications. First, it indicates that the sweep is not restricted to a local population or region. Second, it helps to establish the timing of the sweep, which must have occurred after the out-of-Africa migration of the species, but before the divergence of the European and Asian populations (i.e. 5000–15000 years ago). Third, it suggests that the sweep was not caused by adaptation to a temperate environment per se, as it spans populations from tropical and temperate latitudes. In this respect, the CG9509 example differs from other well-studied polymorphisms in D. melanogaster that show latitudinal clines in frequency and are thought to reflect climatic adaptation [72–74]. Instead, the CG9509 sweep may
be the result of adaptation to human commensalism or agriculture, which is consistent with the inferred role of \textit{CG9509} in detoxification. The sequence variants differing in frequency between the cosmopolitan and sub-Saharan African populations represent candidates for the specific target(s) of selection and future studies that examine their functional effect on \textit{CG9509} expression will help elucidate the molecular mechanism of gene regulatory evolution.

Acknowledgements. We thank John Baines, Sonja Grath, Francesco Paparazzo, Aparup Das, Korbinian von Heuck and John Pool for providing \textit{Drosophila} stocks. We also thank Andreas Massouras and Bart Deplancke for access to polymorphism and eQTL association data for the DGRP lines. Hedwig Gebhart and Hilde Lainer provided excellent technical assistance in the laboratory.

Funding statement. 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.

References


