Supporting Information

Rogers et al. 10.1073/pnas.1006503107

SI Materials and Methods

PCR Amplification Protocol. PCR amplification protocol was 95 °C for 5 min, then 35 cycles of 95 °C for 30 s, 45 °C for 40 s, and 72 °C for 60 s, followed by a 10-min extension at 72 °C, unless otherwise noted. Primer sequences are available in Table S2.

Sequencing the Chimeric Gene. We extracted genomic DNA from each of the *D. melanogaster* strains (listed later) by using the Wizard Genomic DNA purification kit (Promega). The genomic DNA surrounding chimeric gene *Qtzl* was PCR amplified, gel-purified with Quick Spin Gel Purification kit (Qiagen), and sequenced. Primers used in amplification and sequencing, Sequencing F and Sequencing R, were designed to amplify 876 bp of genomic DNA that includes the 5′ and 3′ UTRs of the chimeric gene, as well as some upstream sequence. These sequences were used in Bayesian coalescent inference of effective population size and time to TMRCA.

Probability of Fixation. We amplified and sequenced *Qtzl* in 35 of 35 strains of *D. melanogaster* from a geographically diverse collection. Assuming that presence or absence in each strain is an independent Bernoulli trial, and given a uniform prior distribution on population frequency, we place the lower bound on the frequency of *Qtzl* at 0.920 (95% one-sided CI).

Coalescent Analyses. Fourfold synonymous sites were extracted from the CDS sequences of the chimera *Qtzl*, its eight closest 5′ neighbors, its eight closest 3′ neighbors, and 100 random genes from the nonheterochromatic regions of chromosome 2L. The 68 fourfold sites of *Qtzl* were combined with 53 intronic sites in this analysis. Across these 121 "neutral" sites, we observed a single synonymous polymorphism present in only a single strain. Already, this suggested a recent common ancestor for the sequenced strains of *Qtzl*.

Rigorous estimates of N_e and time TMRCA were obtained through analysis with the program BEAST (version 1.5.2) (1). The HKY85 (2) nucleotide model of substitution was used with equilibrium nucleotide frequencies taken from observed frequencies and transition/transversion ratio κ fixed at 2.389 based on previous work (3). A constant population size coalescent model was assumed with a noninformative Jeffrey prior on N_e. We assumed a neutral mutation rate μ of 5.8×10^{-9} substitutions per site per generation (4) and a generation time τ of 10 generations per year. Estimates of N_e provided by BEAST are scaled in terms of substitutions per site. Converting these estimates requires dividing by μ to get generations and then dividing by 2 to convert generations into individuals assuming a diploid Wright–Fisher model. Estimates of TMRCA provided by BEAST are also scaled in terms of substitutions per site. Converting these estimates to years requires dividing by $\mu\tau$.

A separate analysis was performed using BEAST to estimate the ratio of the rates of nonsynonymous to synonymous substitution ω via a codon based model. In this case, full CDS sequences from *Qtzl* and the same set of 100 random genes were analyzed. Equilibrium codon frequencies were taken from the observed frequencies across all 101 genes. As before, κ was fixed at 2.389. A noninformative Jeffreys prior was assumed for ω. The same constant size coalescent model was used in this analysis.

An independent Markov chain Monte Carlo analysis was perform on each gene. Each chain was run for 11 million steps with the first 1 million steps discarded as "burn-in." Samples were logged every 1,000 steps, providing a total of 10 thousand samples for each gene. Additionally, in a sensitivity analysis, we found our results to be robust to the choice of prior distributions.

Analysis of Selective Sweep. Selective sweeps cause a local reduction in genetic diversity. Kaplan et al. (5) present a series of differential equations that describe the expectation of π for a neutral locus linked to a selected locus that has undergone a recent selective sweep. These equations do not admit closed-form solutions. By using Mathematica (version 7.0), we solved these equations numerically to find the expectation of π moving outward from a selective sweep with selective coefficient s and time of fixation t_f. We fit these expectations to the observed genetic diversity surrounding *Qtzl* on chromosome 2L. We assumed a neutral mutation rate μ of 5.8×10^{-9} substitutions per site per generation (4), a generation time τ of 10 generations per year, an effective population N_e of 1.85×10^6 based on the preceding coalescent analysis, and rate of recombination r of 3.5×10^{-8} events per site per generation, obtained from the *D. melanogaster* Recombination Rate Calculator (version 2.1; http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl) (6). Parameters s and t_f were estimated through least-squares minimization. These estimates assume that regions surrounding *Qtzl* are completely neutral and that demography is simple, i.e., population size has remained constant through time.

Haplotype Structure. We used Sweep 1.0 (http://www.broadinstitute.org/mpg/sweep/index.html; accessed December 2009) to construct haplotype bifurcation diagrams and to calculate EHH (7) for our region of interest (11,945–12,045 kb) using markers from the *Drosophila* DPGP data. Markers were defined every 10th base pair, using only markers with full coverage for all strains.

Comparison of Duplicate Regions. Because reads from duplicate regions do not map uniquely, sequence for the region immediately surrounding *Qtzl* is not available in the standard DPGP release. To determine whether duplicate sites in this region had differentiated, we examined the available sequencing reads from the DPGP project. We downloaded all available raw reads available for 15 Raleigh strains from the National Center for Biotechnology Information database (Accessed August 2009). We used MAQ (Heng Li; http://maq.sourceforge.net/index.shtml) to map all reads to a 5,016-bp template, which included a single copy of from the duplicated region containing *Qtzl* along with and additional 1,105 bp of sequence from *escl*.

RNA Extractions and RT-PCR Assays. We extracted RNA using a standard phenol-chloroform protocol. We pulverized whole or dissected flies in 1 mL of TRIzol (Invitrogen), and incubated at room temperature for 5 min. For whole flies or carcasses, the TRIzol suspensions were centrifuged 10 min to precipitate exoskeleton material. TRIzol suspension was added to 200 μL of chloroform in a phase-lock gel tube (Eppendorf) and agitated for 30 s, then left at room temperature for 3 min. Samples were centrifuged for 10 min at 4 °C. The top, clear phase was removed and added to 500 μL of isopropanol and mixed by inversion for 10 min, then centrifuged for 10 min at 4 °C. Supernatant was removed, and the pellet was washed with isopropanol and centrifuged again for 10 min at 4 °C. Isopropanol was removed, and the pellet was washed with 75% ethanol and centrifuged for 10 min at 4 °C. Ethanol was removed and the pellet was allowed to dry. Nucleic acid was suspended in nuclease free water.

We diluted RNA to a concentration of approximately 20 ng/μL, treated with Turbo DNase (Ambion) according to the standard

protocol, and prepared cDNA from 7 μL of DNase-treated RNA using an oligo-dT 18mer with the SuperScript II system (Promega). To determine expression of the chimeric gene in heads, testes, and carcasses we designed gene-specific primers (*Qtzl* F, *Qtzl* R; *CG12264* F, *CG12264* R; *escl* F, *escl* R) internal to the annotated mRNA for each gene. We amplified transcripts by using 1.5 μL of cDNA in 50 μL PCR reactions. Similar reactions were performed using RNA negative control primers to exclude genomic DNA contamination. Transcript amplification products were visualized on 1% agarose gels. Amplification products from each gene were gel purified from at least one assay and sequenced to ascertain accuracy. To verify the predicted transcript model of *Qtzl*, we amplified, gel-purified, and sequenced the *Qtzl* mRNA using primers designed to amplify a longer segment of the transcript (mRNA F, mRNA R).

Dosage Changes After Duplication. We wanted to determine the relative increase or decrease in the mRNA levels of duplicate genes *Ada1-1*/*Ada1-2* and *CG18787*/*CG18789* relative to singleton orthologues in *D. sechellia*. We mapped all reads from whole-transcriptome RNA-seq experiments SRX016674 and SRX016673 (8) available from the National Center for Biotechnology Information Short Read Archive database (ftp://ftp.ncbi.nlm.nih.gov/; Accessed March 2010) to a single copy template spanning the genes *Ada1-1*, *CG18787*, and *CG12264*. *CG12264* is expressed at high levels in all tissues of *D. melanogaster* and provided a reference point used to normalize expression levels. We excluded a short low-complexity region in *CG18787*, which recruits large numbers of reads that do not perfectly match the reference. The number of reads mapping to each segment of the coding sequence should be indicative of mRNA levels for that gene, although read depth across the gene may vary. For each gene, we randomly chose 10,000 replicates from the distribution of read depths across the mRNA sequence. We then calculated the ratio of read depths for each gene compared with *CG12264* for each replicate. A two-sided Welch *t* test was used to compare the distribution of ratios in *D. sechellia* and *D. melanogaster* to determine significance of expression differences.

Expression from Mutant African Strain. We isolated mRNA and generated cDNA as described earlier from mixed pupae of African strain 191. We amplified and sequenced a portion of the transcript using internal primers *Qtzl* F and *Qtzl* R. No amplification was observed using primers mRNA F and mRNA R. Additionally, we isolated mRNA from the head, testes, and carcass of adult male flies, and amplified using primers *Qtzl* F and *Qtzl* R.

Mapping of the Overexpression Mutant. We obtained an overexpression line of *D. melanogaster* produced by the Drosophila Gene Search Project from Naoyuki Fuse (Kyoto University, Kyoto, Japan). We isolated genomic DNA from whole mutant flies, and amplified sequence adjacent to the *P*-element insertion using primers *P*-element F and *Qtzl* R. No amplification was observed using primers *P*-element F and *CG12264R*. Product was gel purified and sequenced. Sequence clearly aligns to the chimeric gene *Qtzl*.

Characterization of Known Protein Domains. We used the Pfam 23.0 online database (http://pfam.sanger.ac.uk/; Accessed August 2009) to search for protein domains within the chimeric and parental gene sequence. We used BLASTp to compare protein sequence for each gene in the duplicated region against the NCBI nonredundant protein database to identify putative orthologues in other model organisms and to identify known conserved protein domains. We likewise used the TargetP 1.1 Web server (http://www.cbs.dtu.dk/services/TargetP/; Accessed August 2009) to search for known signaling peptides in each gene sequence.

Fly Strains. ***D. melanogaster African Strains (from Daven Presgraves).***

58
81
82
84
125
131
145
159
178
184
191
196

WorldWide strains from the Tucson Stock Center.

14021–0231.04
14021–0231.07
14021–0231.34
14021–0231.40
14021–0231.45
14021–0231.49
14021–0231.51
14021–0231.53
14021–0231.54
14021–0231.56
14021–0231.59
14021–0231.69

Raleigh Strains (from Mike Eisen).

RAL 208
RAL 304
RAL 315
RAL 324
RAL 352
RAL 375
RAL 380
RAL 437
RAL 517
RAL 786
RAL 820

From Sarah B. Kingan.

D. simulans.

14021–0251.007
14021–0251.009
14021–0251.169
14021–0251.176
14021–0251.184
14021–0251.191
14021–0251.194
14021–0251.195
14021–0251.196
14021–0251.197
14021–0251.198
14021–0251.199
14021–0251.216

D. sechellia.

14021–0248.03
14021–0248.05
14021–0248.07
14021–0248.08
14021–0248.12
14021–0248.25

1. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.
2. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174.
3. Bedford T, Hartl DL (2008) Overdispersion of the molecular clock: Temporal variation of gene-specific substitution rates in Drosophila. *Mol Biol Evol* 25:1631–1638.
4. Haag-Liautard C, et al. (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445:82–85.
5. Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. *Genetics* 123:887–899.
6. Singh ND, Arndt PF, Petrov DA (2005) Effect of recombination on patterns of substitution in *Drosophila*. *Genetics* 169:709–722.
7. Sabeti PC, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
8. McManus CJ, et al. (2010) Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res*, in press.

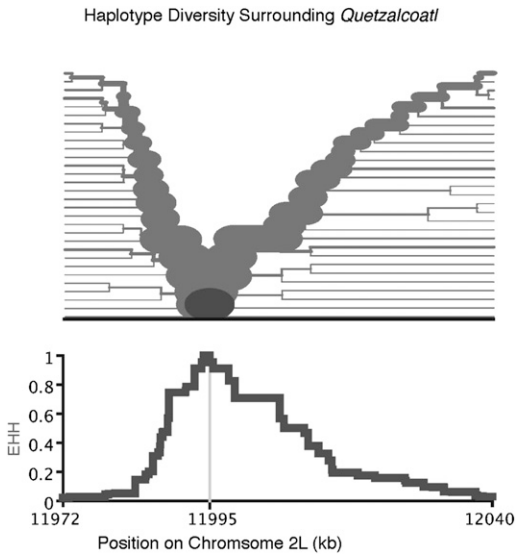

Fig. S1. Haplotype diversity and EHH surrounding *Qtzl*. The region contains a single haplotype, which bifurcates across some 20 kb to the left and 40 kb to the right of the gene. EHH is 1.0 surrounding *Qtzl* and decays outward from the region.

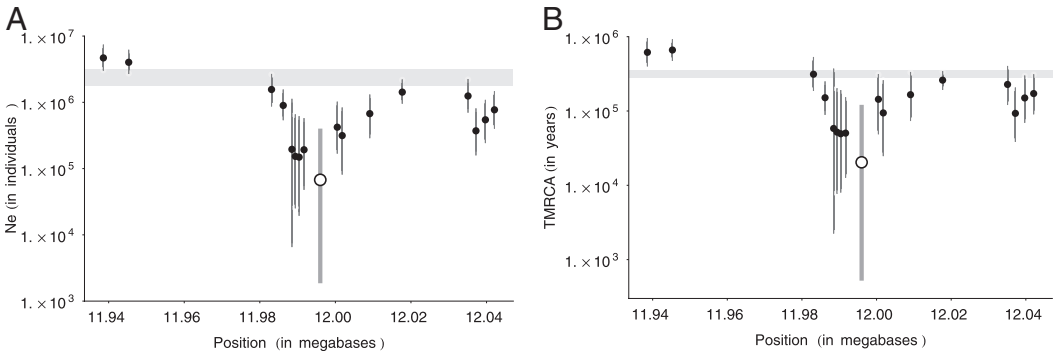

Fig. S2. N_e (*A*) and time to TMRCA (*B*) for *Qtzl* (open circle) and surrounding genes. Chromosomal averages and 95% credible intervals are shown as shaded bands. *Ne* and TMRCA both decrease closer to the chimeric gene, and *Qtzl* displays the most significant departure from the chromosomal average.

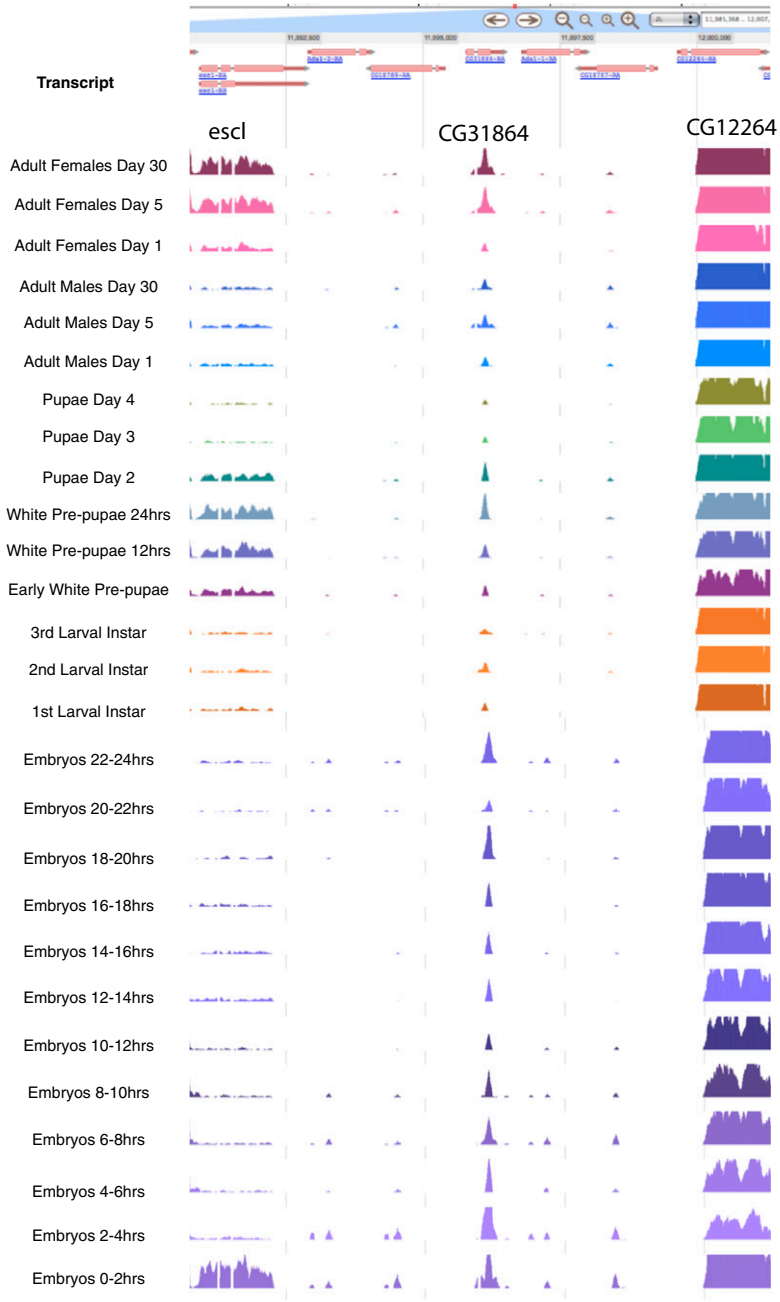

Fig. S3. Transcription sequencing data provided by modENCODE, displaying only uniquely mapping reads for *Qtzl* and surrounding genes. Peak heights are proportional to observed mRNA levels. *Escl* (*Left*) is expressed in early embryos, white prepupae, and in adult females. *Qtzl* (*Center*) is strongly expressed in all embryonic stages, white prepupae, and in adult females. *CG12264* (*Right*) exhibits strong expression at all time points.

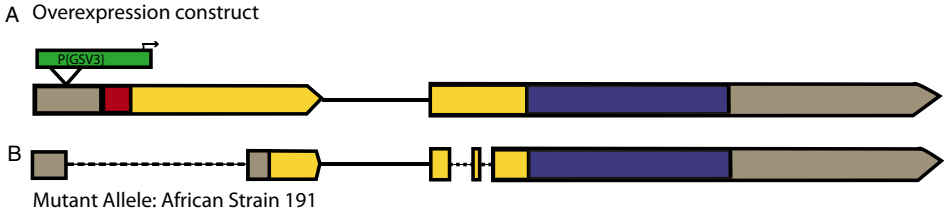

Fig. S4. Structure of two mutant alleles for the chimeric gene *Qtzl*. (*A*) Transgenic *P*-element overexpression Gene Search Vector construct drives overexpression of *Qtzl* from the 3′ end of the construct in the presence of Gal4. In the absence of Gal4, the construct serves as a disruption mutant. (*B*) A naturally occurring allele isolated from Malawi strain 191 carries a series of complex changes. A deletion at the 5′ end just after the transcription start corresponds precisely to the insertion point of the transgenic construct above. Additionally, an indel just after the first intron replaces a 25 bp segment with 17 bp of new sequence. The longest available translation for this allele includes the frameshifted segment inherited from *escl* as well as a short segment from *CG12264*.

Table S1. Population genetic estimates and 95% credible intervals for *Qtzl* and chromosome 2L

Gene/chromosome	π, Substitutions per site	Tajima's *D*	N_e, 10^5 of individuals	TMRCA*, 10^4 y	ω
Quetzalcoatl	0.0004	−2.293	0.676 (0.020–3.73)	2.04 (0.056–11.3)	1.659 (0.239–26.076)
Chromosome 2L average	0.0044 (0.0014–0.0077)	−1.09 (−1.90 to −0.013)	18.5 (17.5–31.5)	30.5 (28.5–35.7)	0.101 (0.084–0.119)
Percentile rank/*P* value	4.5×10^{-4}	1.59×10^{-3}	$P < 10^{-4}$	$P < 10^{-4}$	$P = 0.002$

*Estimates assume a neutral mutation rate of 5.8×10^{-9} substitutions per site per year and 10 generations per year.

Table S2. Primers for sequencing and amplification

Primer	Sequence
Sequencing F	CGCAATTTCGTGAATTCGTA
Sequencing R	GTCGCACACGCAAACTAAAA
Qtzl F	ATGCATTAGCCAATGTGGT
Qtzl R	ACAAATGTTGTGCCCAGTTG
CG12264 F	ACATCCAAAGAATTCCGGG
CG12264 R	CTGACGCCGATTTCGTTATT
escl F	CTACGCAGACCCAGATACCG
escl R	GTCTGTCGGATGGTGGAGTT
RNA negative control F	CTGTCCATTGTCGGTTGTTG
RNA negative control R	GGCCGAGTTGCATAAATGTT
mRNA F	GCATCCAGGCGCAGGTCCTC
mRNA R	CCATCGCCACACGGTGCTGT
P-element F	CTCACTCAGACTCAATACGAC

Other Supporting Information Files

Dataset S1 (TXT)