SUPPLEMENTAL DATA

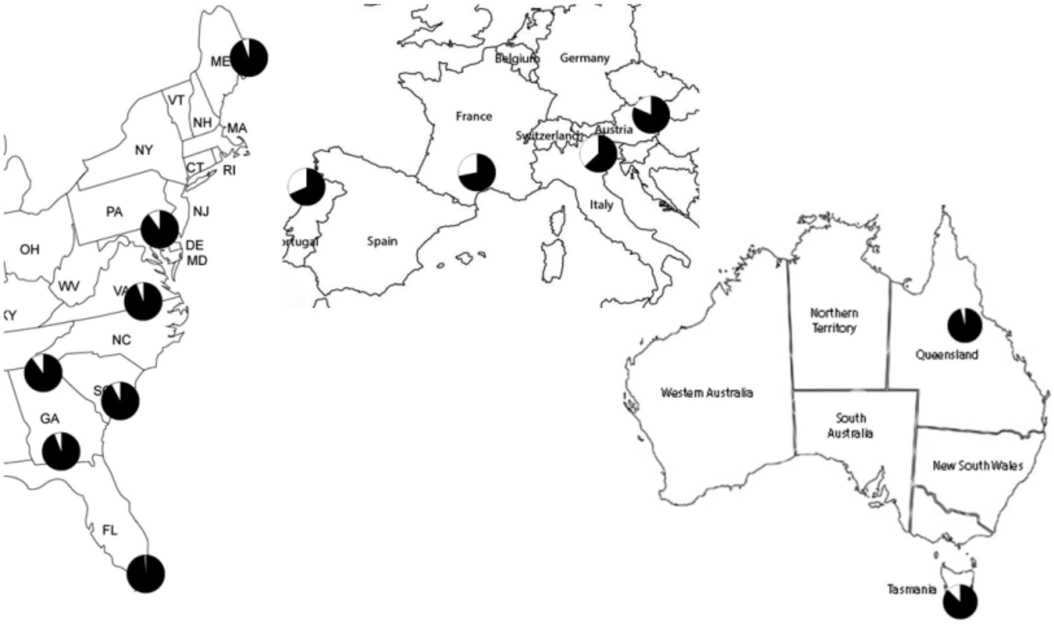

Figure S1. Frequency of Leu (black) and Phe (white) Alleles in *D. melanogaster* Populations. Related to Figure 1. Frequencies were obtained from publicly-available high throughput sequencing data (see Supplemental Experimental Procedures). If the North American data are combined with that from Fry et al. [S1], there is a significant positive correlation between latitude and arcsin-square root transformed frequency of the Phe allele (r = 0.54, N = 16, P = 0.031). Data from Australia are more limited, but collectively support a temperate-tropical difference (Phe allele frequencies in the northern and southern Australian samples of Fry et al. were 0.025 and 0.22, respectively). Moreover, the frequency of the Phe allele in a set of 110 haploid genomes from 22 sub-Saharan African locations was 0.018, similar to its frequency in Florida and Queensland, and substantially lower than the European frequencies.

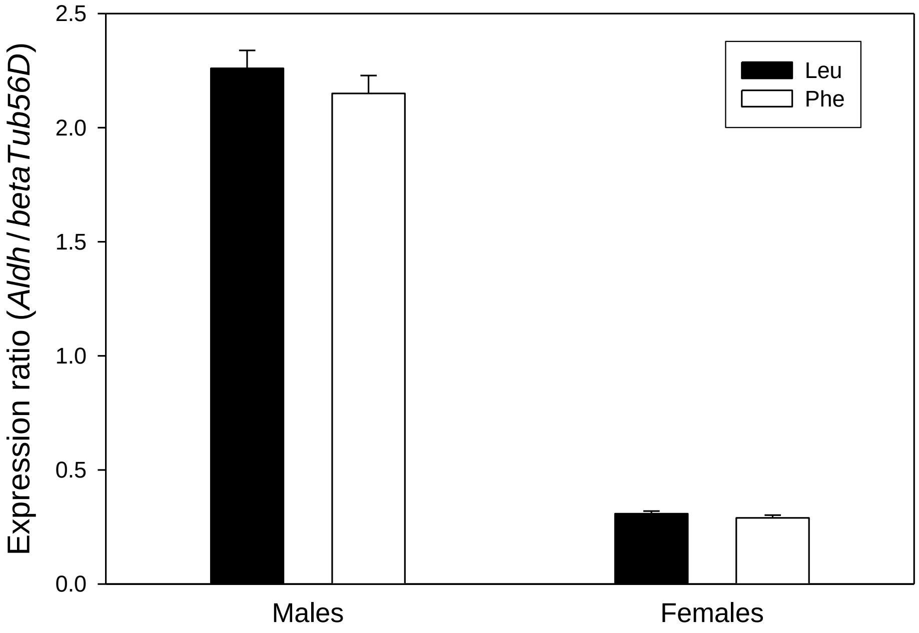

Figure S2. Expression of *Aldh* (± 1 S.E.M) in the insert lines, normalized by the reference gene *betaTub56D*. Related to Figure 2. Differences between genotypes were not significant ($P > 0.3$) in either sex.

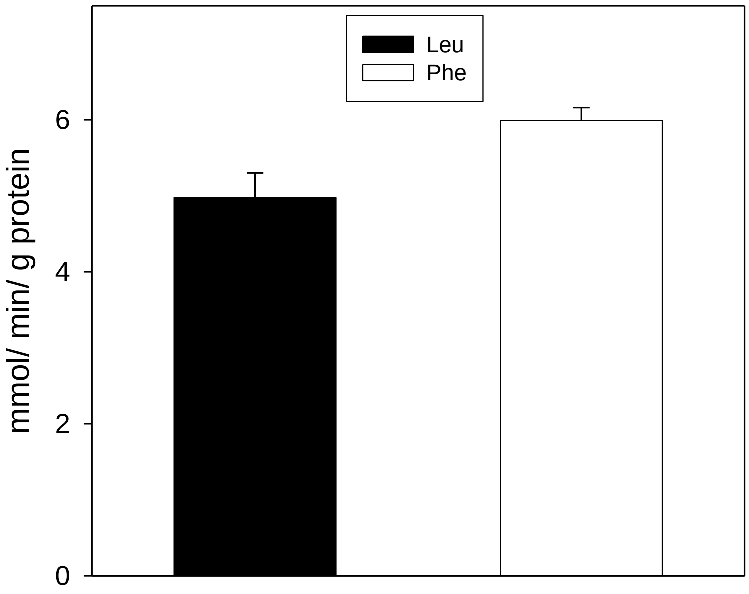

Figure S3. ALDH activity (± 1 S.E.M) of the insert lines with acetaldehyde as substrate. Related to Figure 2. Phe lines had significantly higher activity than Leu lines (*P* < 0.03 one-tailed).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Allele frequencies in natural populations

Illumina short reads of genomic DNA were downloaded from the NCBI Sequence Read Archive. Pooled-seq data, based on population samples of a minimum of 15 isofemale lines, were obtained from the following locations and sources: five North American populations (Florida, Georgia, South Carolina, Pennsylvania, Maine) [S2]; a French and a second Georgia (U.S.A.) population [S3]; populations from Austria and Italy [S4]; two Australian populations [S5]; one Portuguese population [S6]. In addition, genome sequences of 162 Raleigh lines were obtained from the Drosophila Genetic Reference Panel (DGRP) [S7], and sequences of 110 sub-Saharan African lines were obtained from the genome assemblies made available by the Drosophila Population Genomics Project (DPGP2) [S8].

Single and paired end reads were aligned to the reference genome sequence by bwa 0.7.8 [S9]. The SAM alignment files were then converted to BAM files using SAMtools [S10]. Allele frequencies were calculated from the position sorted BAM files.

We used similar methods to align Illumina paired end reads from 270 North American strains of *D. simulans* (NCBI SRA project accession no. PRJNA279205) to the genomic sequence of *D. simulans* Aldh (FlyBase ID GD23600).

Creation of *Aldh* insert lines

We cloned a genomic fragment containing *Aldh*, amplified from the BAC clone BACR20B09 (Berkeley Drosophila Genome Project) [S11], and used ϕC31 integrase mediated recombination to insert it into the fly genome [S12]. The BAC clone was extracted and purified using the BACMAX reagent following the manufacturer's protocol (Epicentre Biotechnologies, Madison, WI). An 8153 bp fragment containing the complete *Aldh* coding sequence along with 2Kb of both upstream and downstream sequence was amplified using the forward and reverse primers 5' GAGGGAGGAAAAGGGTGAAG 3' and 5'TTTAATTATCCTGCCACGCC 3'. The amplified fragment was cloned into the TOPO XL vector (Invitrogen). Because the BAC clone is derived from the reference strain, the amplified fragment had the Leu allele. The Phe allele was then created by site directed mutagenesis. We used the Phusion DNA polymerase (New England Biolabs, Ipswich, MA) to amplify the TOPO XL-*Aldh* construct using the reverse primer- 5'CCTGCGTGCCGGAACTGTGTG 3'- and the forward, mutagenizing primer- 5'GGTAAACACCTACAATGTCTTCGCTGCCCAGGC 3', where the mutation-inducing base is underlined.

The genomic fragments containing the Leu and Phe alleles were excised from the TOPO XL-*Aldh* construct by the restriction enzymes NotI and DpnI and inserted into the pattB vector [S12]. After sequencing the inserts to confirm that they were free of any unwanted mutations, the pattB-*Aldh* constructs were injected into fly embryos containing the *white* marker and an attP receptor site on the third chromosome (cytological band 86F) by Bestgene Inc. (Chino Hills, CA). Transformant flies, recognizable by red eye color, were used to establish lines homozygous for the inserts by crossing to a stock with the third chromosome balancer *TM3*, following standard procedures. Subsequently, the insert-carrying 3rd chromosomes were brought into an *Aldh*-null genetic background by crossing to the balancer stock *w+; $Aldh^{\Delta 17}$ b; TM6C, Sb/H,* where $Aldh^{\Delta 17}$ is a mutant lacking detectable *Aldh* expression [S13] and *b* is the visible marker *black*. We used three transformant lines derived from independent integrations for each allele for our experiments.

Experimental populations and fitness estimates

To measure relative fitness of the Leu and Phe alleles, four experimental populations were established on both normal medium (standard cornmeal-molasses-agar-brewer's yeast recipe) and medium supplemented with 6% ethanol. To prepare the latter, ethanol was added after the medium had cooled below 50ºC in order to minimize evaporation. Each population was initiated from approximately 1200 F_1 individuals from a cross between one Phe line and one Leu line. (For the first three experimental populations, Leu line *i* was crossed to Phe line *i*, where *i* = 1,2,3; for the fourth population, a new combination of two randomly chosen lines was used). Each generation of each population was maintained in 20 shell vials, with approximately 60 flies per vial; adults were allowed to lay eggs for five days and then removed. To establish the next generation, flies that had emerged by the 14th (normal medium) or 21st (ethanol-supplemented medium) day were pooled and redistributed over 20 new vials, using light CO_2 anesthesia. The longer generation time on ethanol-supplemented food was necessary because ethanol slowed development by several days.

To estimate allele frequency changes, DNA was extracted from 23-30 individual F_{10} flies per population and amplified using the forward and reverse primers 5' CCGATGTCCAGGATGATATG 3' and 5' CATATGTACTAGATAGAAATG 3'. This primer pair was designed to amplify the region containing the polymorphism in the inserted *Aldh* locus without amplifying the endogenous *Aldh* locus, because the reverse primer binding site in the latter is split by a naturally segregating insertion (~50bp) relative to the reference sequence. Samples were genotyped either by the PCR-RFLP procedure of Fry et al. [S1], or by sequencing the PCR products.

For each medium type, we estimated the relative fitness values of the three genotypes from the F_{10} allele frequencies by iterating standard equations for allele frequency change, assuming random mating and nine generations of viability selection (F_2-F_{10}; if instead selection had been on fertility, there would have been only eight generations of selection, which would have caused us to slightly underestimate selection coefficients). The unfavored (i.e., minority) homozygote was assumed to have a relative fitness of one, with heterozygotes and favored homozygotes having relative fitness values of Exp(hs) and Exp(s), respectively. Because it is not possible to estimate both h and s from our data, we estimated s for three fixed values of h (0.1, 0.5, 0.9), spanning the range from nearly complete recessivity to nearly complete dominance of the favored allele. For each value of h and each treatment (regular or ethanol-supplemented food), we found the values of s that gave the closest match to the average, lower 95% confidence limit, and upper 95% confidence limit of the final allele frequencies. Stochastic simulations of sets of four small populations, with population sizes ranging from 100-1200, showed that this method produces unbiased estimates of s, and confidence intervals that have the nominal (i.e., 95%) probability of containing the true value.

Aldh expression in insert lines

To determine whether the mutation causing the Leu-Phe substitution affected *Aldh* expression, we measured *Aldh* transcript levels in the six lines by real time PCR. Total RNA was extracted from 15 2-4 day old male flies from each line using the RNAeasy mini kit (Qiagen, Valencia, CA), and cDNA synthesized using the iScript cDNA synthesis kit (Biorad, Hercules, CA). Relative abundance of *Aldh* transcript was measured using the housekeeping gene *βTub56D* as a reference. Real time PCR reactions were performed in an Applied Biosystems 7300 real time PCR system, with three wells per subline, using TaqMan probes (01809880_g1 for *Aldh* and 02362299_u1 for *βTub56D*), following the manufacturer's protocol (Life Technologies, Grand Island, NY). Relative expression of *Aldh* ($2^{\Delta CT}$) did not differ between Phe and Leu lines (Figure S2). Measurements of *Aldh* expression of a set of wild-type lines using the same methods showed that expression of the insert lines was in the wild-type range (higher than that of lines from two locations in Africa, and lower than that of lines from two European locations; M. Chakraborty and J. Fry, unpublished data).

Expression and purification of recombinant ALDH

We cloned *Aldh* cDNA, minus the mitochondrial leader sequence, into the expression vector pMALc2x (New England Biolabs, Ipswich, MA) following the procedure of Rothacker and Ilg [S14]. We used the *Aldh* cDNA clone (GH22814) available from the Drosophila Genomics Research Center (Bloomington, IN) as the source of the *Aldh*^{Leu} cDNA, and a clone kindly shared by Dr. Thomas Ilg as the source of the *Aldh*^{Phe} cDNA. Using a pair of primers designed by Rothacker and Ilg [S14], we cloned the *Aldh* ORF into the BamHI-HindIII restriction sites, located downstream of the open reading frame of the maltose binding protein within the vector. Next, we transformed TB1 cells (New England Biolabs) with the resulting constructs, and induced over-expression of the MBP-ALDH following the manufacturer's protocol. To purify the protein, we ran the crude cellular extract through a 1 ml MBPTrap-HP amylose column (GE Healthcare Life Sciences, Pittsburgh, PA) connected to an automated FPLC machine. The purified protein was eluted using 50mM amylose solution, and stored at –80°C with 43% glycerol and 1 mM DTT. Enzyme concentrations were measured using Lowry reagent (Sigma-Aldrich, St. Louis, MO).

Enzyme kinetics assays

Enzyme assays were carried out at 25°C in cuvettes of 1 cm path length in 1 ml volume of Na_2HPO_4/NaH_2PO_4 buffer (pH 8.5) containing 1 mM DTT and 1 mM NAD+. The reduction rate of NAD+ was measured by following the change in absorbance at 340 nm in an Ultrospec spectrophotometer (GE Healthcare Life Sciences). All substrates were purchased from Sigma and dissolved in DMSO before use. For acetaldehyde, butanal, hexanal, and benzaldehyde, we used the saturating concentration ([S] >> K_M) of 1 mM [S14]. The more reactive substrates trans-2-hexenal and trans-2-octenal were used at a concentration of 1 uM because they inhibited ALDH at 1 mM; this lower concentration is nonetheless likely to be close to saturating [S14]. Each assay was performed in three replicates, with assays for both forms of the enzyme for a specific substrate always being conducted in pairs. Results are expressed as mmol NAD+ reduced per minute per g of enzyme, using the extinction coefficient $\mathcal{E}_{340}$ = 6220 M^{-1} cm^{-1}.

To compare the aldehyde dehydrogenase activity of the insert lines, we ground 20 2-4 day old males in 500 ul of grinding buffer (0.25 M sucrose, 5mM EDTA, 15 mM Triton X-100, 5 mM DTT). Protease inhibitor (Roche Applied Science) was added to the grinding buffer to prevent proteolytic degradation of the enzyme, at the manufacturer's recommended concentration. The extract was kept on ice for 15 minutes and then centrifuged at 16,300 g for 20 minutes, after which the supernatant was transferred to chilled 1.5ml tubes. Protein concentration of a sample of supernatant was measured using a Qubit fluorimeter (Life Technologies), following the manufacturer's protocol. ALDH activity of 100 ul of supernatant with acetaldehyde as a substrate was measured in the same way as for the purified recombinant proteins, except that pyrazole (0.02 M final concentration) was added to the reaction mix to inhibit alcohol dehydrogenase, which would otherwise use NADH to reduce acetaldehyde to ethanol.

Protein structure modeling

Structural models of *D. melanogaster* $ALDH^{Leu}$ and $ALDH^{Phe}$ were constructed by Populus [S15] using the structure of human mitochondrial aldehyde dehydrogenase (ALDH2; PDB id 1O04) and sheep liver class I aldehyde dehydrogenase (PDB id 1BXS) as templates. The model structures of $ALDH^{Leu}$ and $ALDH^{Phe}$ were aligned in Pymol. We measured the volume of the substrate entry channel using FRED Receptor v 2.2.5 (Openeye Scientific, Santa Fe, NM) [S16]. First, potential substrate-binding spaces were searched within the enzyme with molecular probes. Next, using the prior knowledge of the active site from alignments between human ALDH2 and DmALDH, a box was created containing the substrate entry channel. The volume of the channel was then computed using molecular probes.

Hyperoxia and acetaldehyde resistance assays

Hyperoxia resistance was measured by placing 15 0-2 day old male flies in vials with normal medium, with 10 vials per insert line. After allowing flies to recover from anesthesia for one day, the vials were placed inside an airtight plastic container with an inlet tube through which 100% oxygen entered from a tank. Oxygen concentration, as determined by an oxygen meter (Extech Instruments, Waltham, MA), was maintained in the range 90-100%. Flies were transferred to vials with fresh medium after the third day and survival of the flies was recorded after the fifth day. The assay was repeated in a second block using flies derived from a different generation.

To measure acetaldehyde resistance, 20 2-4 day old males or females were placed into food vials. After one day to recover from anesthesia, the flies were mass-transferred to assay vials (4-5 per sex and line), each of which had a ~0.5 g cotton ball at the bottom moistened with 2 ml of a solution containing 2.5% sucrose and 0.85% acetone. The acetone was used to inhibit activity of alcohol dehydrogenase, removing the contribution of this enzyme to detoxification of acetaldehyde [S17]; the concentration used caused no mortality of flies on its own. A second 0.5 g cotton ball was placed into the middle of each vial, trapping the flies between the two cotton balls, and the vial was then sealed with a cork. After 8 hours the corks were briefly removed, and 700 ul of a 2.5% or 3.5% acetaldehyde solution was added to the top of the second cotton ball in vials containing males and females, respectively (vials containing females received 3.5% acetaldehyde because 2.5% induced no mortality). The number of dead flies was counted after the third day. Results for males are shown; results for females were similar (difference between Phe and Leu lines, $P < 0.05$).

Statistical analysis

Enzyme activities were compared between genotypes by two-sample *t*-tests. Acetaldehyde resistance (arcsin square-root transformed proportion surviving) and *Aldh* expression were compared between genotypes by nested analysis of variance, with lines nested within genotypes, using the MIXED procedure in SAS (Cary, NC). For hyperoxia resistance (also arcsin square-root transformed), random effects of block (i.e., assay generation), as well as the genotype × block interactions, were also included (the line × block interaction was dropped after it was found to explain 0% of the variance). For acetaldehyde and hyperoxia resistance, and ALDH activity of the insert lines, one-tailed *P* values are reported, because based on the kinetic results using purified enzyme (Figure 2) there were clear directional expectations in these cases (Leu > Phe for hyperoxia resistance, Phe > Leu for the others). All other *P* values are two-tailed.

Application of Levene model to fitness estimates

We investigated whether environmental heterogeneity could maintain the Phe-Leu polymorphism in a panmictic population, given the fitness estimates in Table 1, using the classic model of Levene [S18]. For simplicity, we assumed that there are two resource types, fruit with no ethanol, and fruit with 6% ethanol, whose relative contributions to the breeding pool of adults are c and $1 - c$, respectively. Given these assumptions, if the relative fitness of the heterozygote on each resource is set to 1, a polymorphism will be maintained if the harmonic mean fitness of each of the two homozygotes is less than 1. Using the point estimates of s in Table 1 for $h = 0.9$, the relative fitness values of Leu and Phe homozygotes in the absence of ethanol are 1.03 and 0.76, respectively; the corresponding values on 6% ethanol are 0.85 and 1.02. The harmonic mean fitness of Leu homozygotes is $(c/1.03 + (1 – \mathrm{c})/0.85)^{-1}$, which will be less than one when $\mathrm{c} < 0.86$. The harmonic mean fitness of Phe homozygotes is $(c/0.76 + (1 – \mathrm{c})/1.02)^{-1}$, which will be less than 1 when $c > 0.05$. Thus a polymorphism will be maintained over a wide range of values of c. (In contrast, similar calculations show that no polymorphism is possible when $h = 0.5$ or 0.1, reflecting the well known result that polymorphism is more likely if the favored allele in each habitat is partly dominant [S19]). These calculations, of course, are based on numerous unrealistic assumptions (e.g., that the fitness estimates in Table 1 are directly applicable to nature), and therefore should be taken with a grain of salt. They serve only to demonstrate that in principle our fitness estimates could result in the maintenance of polymorphism.

Allele age estimation

We estimated the age of the Phe mutation using the haplotype-sharing method of Gandolfo et al. [S20]. *Aldh* sequences with the Phe allele from DGRP (N=9) were aligned, and for each sequence, the length of continuous haplotype sharing with at least one other Phe sequence, on either side of the Phe mutation, was calculated. Base pair distances were converted into recombination distances assuming a map distance of 4cM/1Mb [S21]. We used the correlated genealogy option [S20]; the uncorrelated genealogy option gave a similar point estimate, with narrower confidence limits. Reported estimates have been multiplied by two to account for the lack of recombination in males.

Conservation of residue in Diptera

We used protein-protein BLAST, with the C-terminal 120 amino acids of DmALDH (residues 401-520) as a query, to confirm the presence of leucine at the site homologous to position 479 in all available Dipteran DmALDH orthologues. These were easily identifiable by their minimum 80% sequence identity to DmALDH over the queried region. Genera were *Drosophila* (11 species), *Musca*, *Glossina*, *Bactrocera* (2 species), *Ceratitis*, *Aedes*, *Anopheles* (3 species), and *Culex*.

Search for *Aldh* duplicates

We searched for *Aldh* duplicates in the DGRP and DPGP2 lines using pecnv [S22], Pindel [S23], and CNVnator [S24]. These three programs use different methods (read pair orientation for pecnv, split read mapping for Pindel, and read depth for CNVnator), so a combination of all three provides a more comprehensive list of duplicates than the individual programs alone [S25]. For Pindel and CNVnator, paired end illumina reads were mapped to the release 5.57 *D. melanogaster* reference genome using bwa mem with default parameters. The sam files containing the aligned reads were converted to sorted bam files using SAMtools. The pecnv pipeline uses bwa aln and samtools for read alignment and alignment sorting, respectively [S22]. All programs were run with the default parameters, with the following exceptions: for pecnv, a coverage cutoff of 3 was used to avoid false positives [S22]; Pindel was run using an insert size of 300; for CNVnator, a bin size of 100 was used due to the relatively high sequence coverage of the sample strains.

REFERENCES

S1. Fry, J.D., Donlon, K., and Saweikis, M. (2008). A world-wide polymorphism in *Aldehyde dehydrogenase* in *Drosophila melanogaster*: evidence for selection mediated by dietary ethanol. Evolution *62*, 66-75.

S2. Bergland, A.O., Behrman, E.L., O'Brien, K.R., Schmidt, P.S., and Petrov, D.A. (2014). Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. PLoS Genet. *10*, e1004775.

S3. Bergman, C.M., and Haddrill, P.R. (2015). Strain-specific and pooled genome sequences for populations of *Drosophila melanogaster* from three continents. F1000research *4*, 31.

S4. Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stobe, P., Futschik, A., and Schlötterer, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. PLoS Genet. *9*, e1003534.

S5. Reinhardt, J.A., Kolaczkowski, B., Jones, C.D., Begun, D.J., and Kern, A.D. (2014). Parallel geographic variation in *Drosophila melanogaster*. Genetics *197*, 361-373.

S6. Tobler, R., Franssen, S.U., Kofler, R., Orozco-Terwengel, P., Nolte, V., Hermisson, J., and Schlötterer, C. (2014). Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. Mol. Biol. Evol. *31*, 364-375.

S7. Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The *Drosophila melanogaster* Genetic Reference Panel. Nature *482*, 173-178.

S8. Pool, J.E., Corbett-Detig, R.B., Sugino, R.P., Stevens, K.A., Cardeno, C.M., Crepeau, M.W., Duchen, P., Emerson, J.J., Saelao, P., Begun, D.J., et al. (2012). Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. *8*, 1-24.

S9. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

S10. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

S11. Hoskins, R.A., Nelson, C.R., Berman, B.P., Laverty, T.R., George, R.A., Ciesiolka, L., Naeemuddin, M., Arenson, A.D., Durbin, J., David, R.G., et al. (2000). A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. Science *287*, 2271-2274.

S12. Bischof, J., Maeda, R.K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for *Drosophila* using germ-line-specific ϕC31 integrases. Proc. Natl. Acad. Sci. USA *104*, 3312-3317.

S13. Fry, J.D., and Saweikis, M. (2006). Aldehyde dehydrogenase is essential for both adult and larval ethanol resistance in *Drosophila melanogaster*. Genet. Res. *87*, 87-92.

S14. Rothacker, B., and Ilg, T. (2008). Functional characterization of a *Drosophila melanogaster* succinic semialdehyde dehydrogenase and a non-specific aldehyde dehydrogenase. Insect Biochem. Mol. Biol. *38*, 354-366.

S15. Offman, M.N., Tournier, A.L., and Bates, P.A. (2008). Alternating evolutionary pressure in a genetic algorithm facilitates protein model selection. BMC Struct. Biol. *8*, 34.

S16. Hawkins, P.C.D., Skillman, A.G., Warren, G.L., Ellingson, B.A., and Stahl, M.T. (2010). Conformer Generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J. Chem. Information and Modeling *50*, 572-584.

S17. Barbancho, M. (1992). Effects of dietary ethanol, acetaldehyde, 2-propanol and acetone on the variation of several enzyme activities involved in alcohol metabolism of *Drosophila melanogaster* adults. Insect Biochem. Mol. Biol. *22*, 269-276.

S18. Levene, H. (1953). Genetic equilibrium when more than one ecological niche is available. Am. Nat. *87*, 331-333.

S19. Charlesworth, B., and Charlesworth, D. (2012). Elements of Evolutionary Genetics (Greenwood Village, Colorado: Roberts).

S20. Gandolfo, L.C., Bahlo, M., and Speed, T.P. (2014). Dating rare mutations from small samples with dense marker data. Genetics *197*, 1315-1327.

S21. Comeron, J.M., Ratnappan, R., and Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. *8*, e1002905.

S22. Rogers, R.L., Cridland, J.M., Shao, L., Hu, T.T., Andolfatto, P., and Thornton, K.R. (2014). Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. Mol. Biol. Evol. *31*, 1750-1766.

S23. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865-2871.

S24. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974-984.

S25. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nature Rev. Genet. *12*, 363-375.