Correction: 17 June 2008

www.sciencemag.org/cgi/content/full/1158078/DC1

Supporting Online Material for

Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in *Drosophila melanogaster*

J. J. Emerson,* Margarida Cardoso-Moreira,* Justin O. Borevitz, Manyuan Long

*To whom correspondence should be addressed.
E-mail: jje@uchicago.edu (J.J.E.); mmoreira@uchicago.edu (M.C.M.)

Published 5 June 2008 on *Science* Express

DOI: 10.1126/science.1158078

This PDF file includes:

Materials and Methods
Figs. S1 to S8
Tables S1 and S2
References and Notes

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/1158078/DC1)

Tables S3 to S7

Correction (17 June 2008): Minor formatting changes have been made to the table of contents. Also, in the Materials and Methods (page 20) and in the legend of figure S3 (page 33), potentially misleading terminology regarding “ploidy” has been corrected.

Contents

1 DNA sources **5**

1.1 Laboratory lines 5

1.2 Natural lines 5

2 Microarray data collection and lab protocols **6**

2.1 DNA preparation 6

2.2 Labeling 6

3 Array Platform **7**

3.1 Reannotation of Affymetrix Drosophila genome tiling array 7

3.1.1 Mapping the probes to the genome 7

3.1.2 Annotation of probe properties 7

4 The Modeling Framework **8**

4.1 Definition of Terms 8

4.2 Parameter Estimation 9

4.2.1 Emission Distributions 9

4.2.2 Effects of Mutation on Probe Intensity 10

4.2.3 Transition matrix 11

5 Calling mutations **11**

5.1 Smoothing 11

5.2 Decoding 12

5.3 Definition of genomic context 12

6 Evaluation of CNP dataset quality **13**

6.1 Evaluation of the duplication dataset . . . 13
6.1.1 Duplication calls . . . 13
6.1.2 Duplication boundaries . . . 15
6.1.3 Duplication frequency . . . 15
6.2 Evaluation of the deletions dataset . . . 15
6.2.1 Deletion calls . . . 16
6.2.2 Deletion boundaries . . . 17
6.2.3 Deletion frequency . . . 18
6.3 Estimating the false negative rate . . . 18
6.4 Polarizing duplications . . . 19
6.5 Physical interpretation vs. biological interpretation . . . 22

7 Detecting natural selection **23**
7.1 Population structure and demography . . . 24
7.1.1 Population structure . . . 24
7.1.2 Demography . . . 25
7.2 Estimating the effects of ascertainment bias . . . 25
7.3 Estimating the effects of error . . . 26
7.4 Testing hypotheses of natural selection . . . 29

References **44**

List of Tables

S1 *D. melanogaster* lines used in this study . . . 39
S2 Summary of population structure . . . 40
S3 Caption for genome wide CNP calls . . . 40

S4 Caption for sequenced breakpoints for CNPs 40
S5 Caption for list of genes completely duplicated or deleted 40
S6 Caption for genes used for silent SNPs 41
S7 Caption for full SFS of all CNPs 41

List of Figures

S1 Geographic distribution of *D. melanogaster* populations in Africa 31
S2 Duplication confirmation strategy 32
S3 Types of copy number change 33
S4 SFS for ancestral duplicates 34
S5 Chromosomal distribution of CNPs 35
S6 Summary of population structure 36
S7 Summary of demographic models 37
S8 Site frequency spectra 38

1 DNA sources

1.1 Laboratory lines

Reference and training data is useful in developing a model to infer copy number change using microarrays. By using the relationship between the probe signal within known mutants and information about those probes, including their single copy probe intensity and their GC content, we can predict the behavior of other probes for which we have never observed mutations. To obtain this information, we chose two strains of *D. melanogaster*, one to represent the non-mutant state and one to represent the mutant states, either duplication or deletion. Since the tiling arrays we chose were designed from the *D. melanogaster* genome sequence of a fruitfly line designed by Celera™ (*S1*), we chose that line as the single copy reference line. This line (indicated as stock 2057 in the Bloomington Stock Center Database) is known by the genotype $y^1 oc^{R3.2}; Gr22b^1 Gr22d^1 cn^1 CG33964^{R4.2} bw^1 sp^1; LysC^1 lab^{R4.2} MstProx^1 GstD5^1 Rh6^1$.

To obtain data for the duplicate state, we chose a line from the aberration collection from FlyBase (*S2*) harboring a relatively large (∼200 Kb) segment translocated duplication called $z^1 w^{1118}; Dp(1;2)w^+70h$, also designated as stock 5409 in the Bloomington Stock Center Database. The genotype of this line indicates that the ancestral paralogous locus derives from the X chromosome in the cytological interval 3A7-8;3C2-3. The duplicate copy was inserted into the left arm of chromosome 2 near cytological band 31A3. Notably, the ancestral locus also harbors the w^{1118} mutation, which is a partial deletion, including part of the first exon and the promoter region $5'$ of the white (w) gene.

1.2 Natural lines

We sampled natural lines from diversity center of *D. melanogaster* in Africa (*S3*–*S5*). The lines used, their locations, and their sources are indicated in Table S1. The samples are all sub-Saharan in origin and range from as far northwest as Cameroon and as far southeast as

Zimbabwe, as indicated in Fig. S1.

2 Microarray data collection and lab protocols

The strategy that we employed for detecting quantitative differences in DNA copy number was through DNA-DNA hybridization between a genomic DNA template extracted from natural lines and oligonucleotide DNA probes affixed to Affymetrix™ tiling arrays. In order to obtain the template DNA appropriate for hybridization to the 25 bp probes present on the Affymetrix™ arrays, we extracted genomic DNA from virgin female flies and fragmented it into 50-200 bp fragments using DNase-I.

2.1 DNA preparation

From each line we extracted 20 μg of gDNA from three replicates of 30 flies using the Puregene DNA extraction kit (30 flies) with an additional phenol-chloroform extraction step. For each sample to be hybridized, 10 μg of gDNA was partially fragmented using DNase I, then end labeled with biotin-ddUTP using terminal transferase from Enzo Life Sciences™ following previous protocols developed in the Long laboratory (*S6, S7*).

2.2 Labeling

The biotinilated DNA was then hybridized to Affymetrix™ full genome tiling arrays. Subsequently, the biotin on the chips was made to bind streptavidin, which was then bound by an anti-streptavidin antibody with additional biotin molecules attached. A streptavidin-phycoerythrin conjugate was then bound to these biotins, which emits light when excited by the scanner's laser which was subsequently detected following the standard Affymetrix™ procedure (*S8*).

3 Array Platform

The Affymetrix™ genome tiling array for *D. melanogaster* is built around a 5 μm platform which permits a square grid of 2,560 probes on a side to be placed onto a silicon wafer. Each probe is comprised of a 25 bp oligonucleotide synthesized directly onto the array. The resulting 6,553,600 probes are partitioned into a number of categories, mainly comprised of pairs of probes whose sequences uniquely match a stretch of genome at either 100% identity or with a single mismatch at the thirteenth basepair. Each pair is situated in adjacent grid positions, with the perfect match probes forming even rows and the mismatch probes comprising the odd rows.

3.1 Reannotation of Affymetrix™ Drosophila genome tiling array

3.1.1 Mapping the probes to the genome

Because the original microarray design was finished before release 4 of the genome was published, reannotation was undertaken. In order to map the probe positions to the genome, all 6,553,600 probes were used as queries against the *D. melanogaster* release 4 (Flybase) genome with Megablast (*S9*) with a word size of 12. This type of search guarantees successful location of homology maintaining at least 15 (word size + 3, see S9) consecutive perfect matches. As a consequence, the search results are reliable only for location of perfect matches to the genome. Mismatches present in any of the 5 middle basepairs of a putative match to the probe would result in an alignment without the required 15 consecutive identities for a guaranteed hit, though some hits are occasionally identified.

3.1.2 Annotation of probe properties

Each probe binds labeled DNA, which then emits light detectable by a scanner. It is not necessarily true that the template DNA bound to the probes has the same sequence as the probe at the coordinates being measured by the scanner. In fact, some probes have the property that they

bind DNA very non-specifically while others are quite specific. The GC content is an important factor in determining the extent of this cross hybridization in addition to the strength of normal hybridization.

4 The Modeling Framework

Due to both heterogeneity in the composition of the probes necessary to interrogate different parts of the genome and due to experimental variation in microarray data, inferences upon single probes in isolation are quite noisy. With replication however, not only can Single Feature Polymorphisms (SFPs) be identified (*S10*), so can CNPs. Because the CNPs span multiple probes, it is desirable to utilize spatial information to improve the quality of inferences drawn from probe data. To address these issues, we have chosen to use Hidden Markov Models (HMMs) to analyze the array data (*S7, S11, S12*). HMMs have the following advantages:

1. spatial information from surrounding probes contributes to the inferences at particular positions as a result of the Markov property;

2. the model parameters can be interpreted as the probability by which mutation states switch between each other, which is informative about rates and sizes;

3. the output of the model permits straightforward decoding schemes that can identify regions of the genome likely involved in copy number variation.

4.1 Definition of Terms

We first consider the genetic unit, in this case a chromosome, which is a sequence T positions long. Each position is indexed by t such that $t \in [1, 2, \ldots, T]$. Each chromosome is associated with R replicate observations, which can be used to infer the mutational state π at position t. In

this framework, our goal is to infer the mutational state sequence $\Pi = \pi_1, \pi_2, \ldots, \pi_T$ with an HMM based on replicate sequences of array hybridization data, X. From the perspective of the model, the states are said to be hidden, as they must be inferred from data that does not directly reveal them. In the model, there are N possible hidden states, indexed variously by i, j, k. As a consequence, $i, j, k \in [1, 2, \ldots, N]$. The states at positions $t-1, t, t+1$ are described by the following: $\pi_{t-1} = i, \pi_t = j, \pi_{t+1} = k$. We have adopted the convention that the state at positions $(t-1, t, t+1)$ are indexed by (i, j, k) respectively (*S13*).

The observed data X were organized in a matrix of quantile normalized (*S14*) and log transformed intensity values representing all chromosome positions replicated R-fold. Thus, as data, we have access to an $(R \times T)$ matrix of observations for each line, where R is the number of replicate chips. From this matrix, we compute a sequence of means of replicate sets along the entire chromosome, $\overline{X} = \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_T$. We also computed a sequence of standard deviations for the mean, $S = s_1, s_2, \ldots, s_T$. For each line, the probes exhibiting the highest 2.5% standard deviations were dropped from the matrix. To guard against artificially estimating a very low probe standard deviation due to few replicate observations, a small constant s_0 was added to each standard deviation as a baseline (*S15*). See the Section 5.1 for details.

4.2 Parameter Estimation

4.2.1 Emission Distributions

There are two classes of parameters in the HMMs we used. The first, or the emissions matrix E, is an $(N \times T)$ matrix of probability functions. The form of the functions depends on the problem being modeled. Because we formulated our question as the difference in log intensity between a sample from a reference strain and a sample from a natural strain, the probability functions were continuous and can be described by statistics that follow t-distributions. For each state, we obtained a $(1 \times T)$ matrix of t-statistics of the following form:

$$T_s = \frac{(\overline{X}_1 - \overline{X}_2) - (M_{Mutant} - M_{Reference})}{\sqrt{\left[\frac{(R_1-1)S_1^2+(R_2-1)S_2^2}{R_1+R_2-2}\right]\left(\frac{R_1+R_2}{R_1R_2}\right)}} \sim t_{R_1+R_2-2} \quad (1)$$

where the subscript 1 denotes the natural line and the subscript 2 denotes the reference line. The upper case of the statistics $T_s, \overline{X}, S$ indicate that they are the full chromosomal sequence of statistics for each replicate set for natural and reference lines. Of all of the quantities in the equation above, only the $(M_{Mutant} - M_{Reference})$ term (i.e. ΔM) cannot be calculated directly from the data. The expected log difference (for ΔM) must be estimated either from a training dataset or from the data using an EM, MCMC, or similar approach. This issue will be discussed in the following section. Except for estimating the effect of mutation, or ΔM, the emissions probabilities is completely specified as soon as the data from the natural and reference lines are collected. In the case of single copy state, it can reasonably be said that $\Delta M = 0$, so in fact, all of the probabilities for the single copy state are specified before estimation of ΔM for the mutant states and follow a simple t-distribution with $R_1 + R_2 - 2$ degrees of freedom.

4.2.2 Effects of Mutation on Probe Intensity

In order to estimate the effect of duplication or deletion on the intensity of probes, we employed a strategy using the intensity measured for known states (single copy, duplicate, and deleted) in order to train a model that predicts the response of any probe in the genome. Because the individual properties of each probe were very important in determining the signal strength and the change in signal caused by mutation, factors that predict these variables are important to identify. Since the most important determinant of change in signal strength is likely to be how strongly the labeled genomic DNA anneals to the probes on the chip, we have identified two important variables to predict this quantity. First, we estimated the strength of hybridization of each probe directly by measuring the hybridization intensity of the known single copy line. In

order to estimate the affinity of hybridization, we used GC content as a proxy. We then used these variables as predictors of intensity of mutant probes in a linear regression. The regression is of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

where y is the intensity measured from probes known to be either duplicated or deleted, x_1 is the intensity of those probes in the single copy genome and x_2 is the GC content of the probe and the β terms are regression coefficients.

4.2.3 Transition matrix

The transition matrix A of an HMM describes the probability of leaving one state and entering another, and determines the expected sizes of copy number variation states. In order to estimate transition matrices for the genome which allow for regionally high or low copy number variation, we fit the maximum likelihood parameters for A in overlapping sliding windows across the genome using the Baum-Welch algorithm (*S11, S12, S16*). The windows were chosen to be 700 probes long, with the tiled central portion being 400 probes and the overlap of each edge being 150 probes. We performed the estimations such that the posteriors from the middle of each window was contributed to the final inference for posteriors, while the overlapping edges were used to avoid edge effects of the HMM. Thus all middle regions were tiled end-to-end across the genome, overlapping by the edges of the windows.

5 Calling mutations

5.1 Smoothing

In order to reduce false positive rate, we tuned the smoothing parameter (which we here call s_0) which we added to the probe standard deviations by an amount sufficient reduce the noise level in the posterior calls. We measured noise as the proportion of probes on a chromosome falling

between a posterior probability of 0.05 and 0.1. For duplicates, we chose s_0 such that the noise was 0.18%. For deletions, we chose s_0 such that the noise was 0.03%.

5.2 Decoding

Upon obtaining posterior probabilities, we called regions exhibiting runs of high posterior probabilities as mutations. To do so, we require the run of probes satisfy the following criteria:

1. the regions exhibiting runs of probes exceeding a threshold of 0.4 posterior probability are first identified;

2. these mutation calls are joined if the distance between them is less than 3 Kb.

5.3 Definition of genomic context

The dataset of CNPs was divided into four mutually exclusive categories according to the genomic annotation of the region mutated. The four categories are:

1. intergenic CNPs;

2. intronic CNPs;

3. exonic CNPs;

4. CNPs encompassing complete genes.

For the following description, we reference release 4.3 of the *D. melanogaster* genome. A CNP was classified as intergenic if it did not overlap a known gene structure, including both protein-coding and non-coding genes. A CNP was classified as intronic if it was contained entirely within an annotated intron in all known transcripts of that gene. If a CNP overlapped sequence that is sometimes transcribed and sometimes removed by alternative splicing, it was

considered exonic and not intronic. Thus, the exonic class includes all CNPs that overlap with any exonic sequence, though they need not overlap exon sequence exclusively. For example, a CNP overlapping part of an exon and part of an intron was classified as exonic. On the contrary, the intronic and intergenic categories required that the CNP falls entirely within intronic and intergenic sequences, respectively, whereas the exonic category could also overlap both intergenic and intronic sequences in addition to exonic sequence. CNPs that contain complete gene structures were classified seperately from the exonic category. Non-coding and coding exons were both considered as exonic. The number of CNPs overlapping exclusively non-coding exons was very small since the majority of CNPs in our dataset overlapped non-coding exons also overlapped coding sequence.

6 Evaluation of CNP dataset quality

We evaluated the quality of the CNP calls made by our model by conducting PCR-based assays on a subset of mutations. Here, we discuss in greater detail the results of the PCR assays and their implications for the overall evaluation of our map of CNPs. We also discuss how well our model can predict the true boundaries of a mutation event and how well the model estimates the true frequency of a CNP in the populations studied.

6.1 Evaluation of the duplication dataset

6.1.1 Duplication calls

In order to confirm predicted duplications, we limited ourselves to the most common mode of duplication, tandem duplication. Our assay relied on the proximity of redundant copies created by tandem duplication. The assay requires the design of two divergent PCR primers located within the predicted boundaries of the putative duplication. With respect to a genome containing only a single copy of the focal region, the primers should be present only in a divergent

configuration. As a result, no amplification should be possible. In the event of a tandem duplication (Fig. S2), the primer configuration changes allowed the primers to become convergent, making amplification possible. One practical limitation of this assay is based on the ability of Taq polymerase to amplify large regions. Due to the occasional presence of spacers of unknown length between tandem copies or if there was a large underestimation of the true size of a duplication, the PCR products were sometimes too large to be amplified, resulting in false negatives. Furthermore, dispersed duplications, such as translocated duplications and retroposed duplicates, could not be confirmed by this method, providing another source of false negatives. We accepted the confirmations as ascertained via PCR as true and therefore called such hybridization signals false positives. As a result, the false positive rate inferred from these experiments was an over-estimate of the true rate, making our inferences conservative.

Another caveat is that this assay imposes a limit on the size of the duplications we can confirm. Because there is some uncertainty associated with the exact location of the boundaries of the duplications, the pair of divergent primers has to be designed well within the predicted limits. One important consequence is that the size of the duplications tested was higher than that of the entire dataset of duplications predicted by our model. The mean size of the duplications present in the confirmation set was 5,031bp (median 3,332bp) while the mean size of the duplications present in the entire dataset was 1,149bp (median 367 bp).

We performed PCR assays on a set of 74 duplications and we obtained a positive confirmation for 64 of them (86%). We then asked if there were any differences between the set of confirmed duplications and the set of mutations that we considered to be false positives (a conservative estimate given the first caveat of this approach). We compared both sets in terms of size, predicted frequency in the populations and genomic context. We found no significant differences between the two sets for all of the above comparisons ($P > 0.05$). There were no size differences between the set of duplications with a positive PCR result and those with

a negative. There were also no differences in terms of the genomic context those duplications overlap, i.e. intergenic, intronic or exonic regions. We included in our confirmation set duplications predicted to be present in only one population (singletons) and duplications predicted to be present in at least two populations (non-singletons). There were 7 duplications predicted to be in at least two populations that we were unable to confirm. That the model predicts the same mutation at least two times independently should add confidence in that call. Hence, we could interpret the absence of confirmation of a non-singleton mutation as suggestive that our PCR assay is indeed conservative and that a fraction of the mutations we were unable to confirm may actually correspond to true mutations.

6.1.2 Duplication boundaries

We sequenced the breakpoints of 48 of the 64 confirmed duplications in order to determine how well our model predicts the boundaries of the mutations. The original data comparing the predicted and experimentally determined boundaries can be found in Table S4. The resolution of the model was limited by the probe size, which is 25bp. We calculated for each boundary (48 x 2 = 96 breakpoints) the difference between the model prediction and the experimental determination. The median absolute difference is 98bp and the mean 738bp. We overestimated the limits of duplications 69% of the times (66/96 breakpoints).

6.1.3 Duplication frequency

We screened by PCR a subset of the confirmed duplications in all 15 natural populations. The model predicted the existence of 48 PCR bands and our screening revealed a total of 57. There was a complete agreement in terms of the lines predicted to have duplication and those that were confirmed to have it. Our model failed to identify 16% of all duplications.

6.2 Evaluation of the deletions dataset

6.2.1 Deletion calls

For this assay, we designed two convergent PCR primers located outside the predicted boundaries of the putative deletions. Since the comparison between a line containing a polymorphic deletion and a line lacking that deletion leads to a PCR fragment length polymorphism, this experiment is usually sufficient to confirm a deletion. At times, no fragment length polymorphism was detected in comparing positive control and the diagnostic experiment. Subsequent sequencing has sometimes shown that when this happens, the physical model is not wrong, but rather the biological interpretation is inappropriate. In these cases, the actual probes predicted by the model to be absent were in fact absent from the band, though other sequence was present instead.We sequenced all bands indicative of deletion and also most bands indicative of absence of deletion (see discussion below). Out of a set of 81 predicted deletions we confirmed 43 (53%). This corresponds to a very high rate of false positives, 47%. We then investigated the differences between the sets of confirmed and false positive deletions in terms of size, predicted frequency and genomic context.

False positive deletions were significantly smaller than true deletions being covered by a smaller number of probes (Wilcoxon rank sum test, P=0.022). This difference disappears if we considered only those mutations predicted to be larger than 200bp. Of the 81 deletions tested, 42 were predicted to be singletons while the remaining 39 were predicted to be non-singletons. We compared the sets of confirmed and false positive deletions in terms of the proportion of singletons and non-singletons. Surprisingly, we found that the set of false positives was enriched with non-singletons (Fisher's exact test, P=0.0146). While the assay used to confirm duplications only gave us a conservative estimate of the rate of false positives, the assay for deletions allowed us to determine unequivocally the true false positive rate. As discussed for the duplication confirmation set, when the model predicts the presence of a given mutation at least two times independently that gives us more confidence that the mutation is real. As

a consequence, the enrichment of non-singletons in the set of false positives was puzzling. We examined in more detail the sequence data collected from the set of false positives. We found that for approximately half of the false positive deletions there were either SNPs (Single Nucleotide Polymorphism) and/or small indels (insertions/deletions) that overlapped with the probes present in the array.

Since our probes were only 25bp long, a SNP or an indel impact the probe signal in a manner similar to a deletion. If the sequence variation present in a haplotype mimics the hybridization intensities characteristic of a deletion, all lines that have that same haplotype will consistently produce the same deletion prediction (*S10*). As a consequence, several independent predictions, while perhaps indicative of a mutation of some sort, is not necessarily exclusively evidence for presence of a deletion. Given the effect of sequence variation in the predictions made by our model, we investigated if such variation was also introducing a bias in terms of the genomic context of the region mutated. This might be relevant if, for example, SNPs were disproportionately more or less common compared to CNPs in certain regions compared to others. Therefore, we conducted contingency analysis on the number of deletions confirmed as true and those confirmed as false in relation to genomic context. We performed two types of genomic context classification. In the first we divided deletions into those overlapping exonic sequence and those overlapping non-exonic sequence. In the second we further divided those deletions that overlap non-exonic sequence into deletions overlapping intergenic sequence and those overlapping intronic sequence. Contingency analyses performed with the two classifications revealed that the false positive group was not enriched for any type of genomic context. There were not significantly more deletions overlapping exonic sequence in the confirmed set than in the set of false positives.

6.2.2 Deletion boundaries

We collected sequence data for the breakpoints of 38 deletions. The goal was in order to determine how well our model predicts the limits of deletions. The original data can be found in Supplementary Table S4. Again, as was the case for duplications, the resolution of the model was limited by the probe size of 25bp. We calculated for each boundary the difference between the model prediction and the experimental determination. The median absolute difference is 32bp and the mean 369bp. From this we found we extend the breakpoint for deletions too far 50% of the times (38/76 breakpoints).

6.2.3 Deletion frequency

We screened by PCR 20 of the confirmed deletions in all 15 natural populations. The model predicted the existence of 48 PCR bands. Our model does not differentiate between deletions homozygous for a population or heterozygous. Moreover, our model was trained with hybridization data collected from homozygous deletions so it will preferentially identify these. The PCR screenings revealed that a subset of the deletions screened were heterozygous in some populations. If only homozygous deletions are considered the model failed to identify 9 deletions (16%). If we also consider heterozygous deletions then the model failed to identify 23 of the 71 bands present in the screening (32%).

6.3 Estimating the false negative rate

In order to estimate the false negative rate, we screened CNPs predicted by our model in all lines, whether or not it was predicted in that line. We calculated the false negative rate to be at least 16% for duplications and homozygous deletions and 32% for heterozygous deletions. We confirmed duplications calls for many values of the smoothing parameter (Section 5.1). The model here described is the one that best recapitulates the experimental data. However, this model failed to predict several CNPs that were predicted by other models. These calls have

also been figured into the error rates. In total we looked at 105 putative duplications of which 80 were experimentally verified. This translated into 16 verified duplications that our model failed to identify. For deletions we looked at a total of 167 putative mutations of which 70 were confirmed to be real deletions. This translates into 27 verified deletions that appear as false negatives in our dataset. Even though these numbers only provide us with a very biased rough -estimate of the false negative rate they suggest that the latter can be potentially very high. Hence, the numbers presented for the amount of copy-number variation in the *D. melanogaster* genome despite the high rates of false positives (especially for deletions) may actually correspond to a considerable underestimation of the true number of mutations.

As a final note, we reiterate that small probes (like the ones used in this study) are strongly affected by variation at the sequence level. In our study this seems to be particularly true for deletions where sequence variation decreased the hybridization intensities in a manner similar to that of actual deletions. This motivates the development of models that are able to take into account sequence variation. This observation also cautions against concluding that non-singleton mutations are verified CNPs, because sequence variation can produce other polymorphisms that are not CNPs yet still give array signatures largely consistent with copy number polymorphism.

6.4 Polarizing duplications

Because tiling array probes are designed by Affymetrix to avoid redundancy in the reference genome, they are all single-copy. In fact, we manually annotated the probes according to Release 4 of the genome assembly (Section 3.1), removing a small handful of probes that were redundant under release 4 of the annotation, but were thought to be unique when the chip was designed. Thus, we were forced by the platform to limit our inferences on copy number changes to those with variation in non-redundant portions of the reference genome. In fact, this limitation introduces a strong ascertainment bias, discussion of which is treated in Section 7.2.

Here we focused on polarizing the CNPs we were capable of ascertaining. In this context, our use of the term "deletion" refers to a change in copy number from a diploid copy-number of 2 in the ancestor to a diploid copy-number of 0 (Fig. S3). Our use of "duplication" in turn refers to a change from 2 to 4 (Fig. S3). A third important process in copy number evolution is the loss of ancestral duplicates, such as occurs by the deletion of an entire locus that was previously duplicated, or a change in diploid copy-number from 4 back to 2 (Fig. S3). Confusingly, this process is also usually referred to by the generic term "deletion".

To remove the ambiguity of these two terms for deletion, we instead refer to "unique deletions" and "redundant deletions". As a consequence of these clarifications, "unique deletions" (change from 2 to 0) and "unique duplications" (change from 2 to 4) are unambiguously distinguishable from each other. Our study did not attempt to explicitly treat "redundant deletions" (change from 4 to 2), although these mutations may be a confounded with "unique duplications" (change from 2 to 4), at least in principle. Such multiple hit scenarios (an older duplication followed by redundant deletion), while certainly important to the long-term fate of duplications, are much less important on the timescale of polymorphism. Indeed, we found empirically that such multiple hits are very rare if not entirely absent in our dataset, even for timeframes extending as far as the *D. melanogaster* and *D. yakuba* split.

To demonstrate this, we first determined whether any of our mutations have evidence for being ancestral duplicates. We used blastz and axtChain software (*S17*) to screen our entire duplication dataset for redundancy in the *D. simulans* and *D. yakuba* genomes. Our methodology requires the following three criteria be met for any duplicate to be considered ancestral:

1. At least two hits are present in either *D. simulans* or *D. yakuba* (i.e. paralogy exists);

2. Paralogs are aligned over at least 40% of their sequences (i.e. the hits aren't extremely short);

3. Paralogs must show 70% identity in the aligned region (i.e., at least some probes remain between aligned regions).

We found that at most 109 duplications show such evidence for the entire dataset (80 for the 10 lines from population 1), most of which were sufficiently divergent to ensure that they don't share more than a handful of array probes preserved between them at most. In fact, 70% identity was a very low threshold for polymorphic duplications, and would identify many paralogs that exceed the age possible for polymorphisms. However, in order to be conservative, we allowed this relaxed threshold. Making the threshold more stringent greatly reduces this number, ameliorating the potential bias from minor to nearly absent. Since this number was so low, for the purposes of SFS analyses, we have excluded these mutations in a conservative attempt to ensure that we took all possible precautions to avoid the potential bias we describe above.

However, examining the SFS of these mutations definitively rejects the hypothesis that a substantial proportion of them are "redundant deletions" instead of unique duplications. Fig. S4 shows the full SFS for all putative "redundant deletions". Notably, the largest frequency class was the singleton class, while the 9-tupleton class contained no observations. If a substantial proportion of these mutations were redundant deletions, then we would expect to have mis-polarized that subset, because the derived mutation would be the one-copy version not the two-copy version. If that were true, then the majority of mis-polarized mutations should have been singletons or doubletons that were mistakenly put into the 9-tupleton or 8-tupleton classes, respectively. Instead, we observed only a single mutation in a combination of both of those classes combined, an observation that is not compatible with ancestral duplication. Under these circumstances, one cannot even assert that the relaxed threshold for paralogy (70% identity) is responsible for swamping out the signature of putative redundant deletions, as there appears to be no signature of mis-polarization at all. If mis-polarized mutants were present, there was

evidence for at most only one doubleton and no singletons. Indeed, this SFS evidence strongly rejects the "redundant deletion" hypothesis for our dataset.

6.5 Physical interpretation vs. biological interpretation

It is important to note that many genuine false positives from the perspective of copy number variation will actually be a result of our imposition of a biological interpretation on a model derived from physical data. For example, we usually describe the duplicate state probability as the chance that a given probe is drawn from a duplicated chromosomal region. However, a more accurate description would be, "Relative to the other $N - 1$ states, the hybridization intensity of light observed from probe t is consistent with the intensity of light expected from known state i with probability $P_{t,i}$". This description highlights a few limitations of the simple model we use. First, the posterior probability is always relative to the other states included in the full model. For example, if a region were triplicated, the best state in our model to describe such a mutation would be a two-fold increase (i.e. duplication) although the actual change was three-fold. Thus, we expect there to be some results that are false positives biologically, even though they are actually true positives with respect to a physical interpretation of the model. The most compelling example of confounding physical interpretation with biological interpretation involves regions in the genome that exhibit unusual variability at the nucleotide level rather than in copy-number. Some relatively small chromosomal regions in one individual from a population can exhibit little to no homology to the same region in other individuals, but no apparent loss of DNA is actually observed. As a result, the sequence in such regions no longer contains the DNA oligomers comprising the probes in the array. In such cases, a mutation or series of mutations has clearly occurred, though the PCR assay indicates that the size of the region has not in fact decreased.

Such cases, while not relevant to our particular biological questions, are still important to

acknowledge. First, this limitation is important to consider when evaluating the results of the method. Second, acknowledging that many biological false positives are 'real' in at least a physical sense lends credence to the statistical performance of the model. On the one hand, it is unfortunate that the most convenient method to detect copy number variation also detects changes that are unrelated to copy-number changes at the locus of interest. In another sense however, it is comforting to know that, given that these confounding factors do exist, the model is capable of identifying them. Third and finally, acknowledging these limitations motivates modifications of the model. In some cases, it might actually be possible to refine the model such that we can identify such confounding factors. Of course doing so would require there to be *some* observable difference in the size, frequency, or intensity of such confounding mutations. In any case, it is clear that hybridization signals allow many biological variants to be distinguished, both from CNPs and otherwise.

7 Detecting natural selection

Because it has been demonstrated that various populations of *D. melanogaster* have experienced demographic histories sufficient to influence the site frequency spectrum (SFS) (*S3, S18–S21*), it is conceivable that our sample too exhibits a skew towards rare variants in the SFS of all genomic regions with respect to the standard neutral model (SNM), even in the absence of natural selection. As a result, standard population genetics approaches with Tajima's D (*S22*) may show evidence for violation of the SNM even in the absence natural selection. Moreover, ascertainment biases and errors in calling mutations (false positives and false negatives) will lead to further violations ofthe assumptions of the SNM. In order to correct for these effects, we must estimate the influence of population structure, demography, bias and error on our data and determine what inferences can be made after such forces are accounted for.

7.1 Population structure and demography

In order to estimate the effects of population structure and demography, we collected a set of 600 synonymous SNPs, sites thought to be evolving under the least constraint. These SNPs were collected from 46 loci located in all major chromosome arms and away from pericentromeric regions (Fig. S5). These loci were chosen from two previous studies that also used silent SNPs as neutral variants (S23, S24 & Table S6). Each locus consisted of a 650-750 bp fragment that was amplified in a single male in all 15 lines used in this study. Contig assembly was performed with Phred Phrap Consed (http://www.phrap.org/phredphrapconsed.html) (*S25–S27*). *D. simulans* or *D. sechellia* orthologous sequences were used as an outgroup and alignments of all sequences were done with ClustalW. Homozygous and heterozygous SNPs were identified with PolyPhred (`http://droog.mbt.washington.edu/PolyPhred.html`; S28).

7.1.1 Population structure

We detected population structure by applying the software STRUCTURE (*S29*) to our silent SNPs in our 15 samples. The results from STRUCURE indicate the presence of two populations among our 15 individuals. Of our 15 lines, 10 individuals were derived almost entirely from population 1 showing nearly no admixture with population 2, 3 were derived almost entirely from population 2 showing nearly no admixture with population 1, while the genomes of two individuals were drawn predominately from one population or the other, but with some evidence for low levels of admixture between the two (Fig. S6). Notably, models positing either only one population or more than two populations were many orders of magnitude less likely than the model for two populations [1 population: $\ln(L_2/L_1) > 47$; 3 populations: $\ln(L_2/L_3) > 1,847$), Table S6]. As a result, we rejected the hypothesis of 1 population in favor of a hypothesis of 2 populations. In order to eliminate the signature of population structure from our subsequent SFS analyses, we conducted all SFS analyses on the subset of original

sample comprised of 10 individuals from population 1.

7.1.2 Demography

In order to account for demography, we applied the procedure of S30 to the synonymous SNPs among 10 individuals from population 1 above to estimate the demographic parameters of two classes of models; a 2-epoch model and a 3-epoch model (Fig. S7).

For the 2-epoch model, we explored the likelihood surface for τ and ν. The parameter τ is the time since the previous epoch in terms of the number of generations scaled by $2N_{current}$. The parameter ν is the relative change in population sizes $\left(\frac{N_{past}}{N_{current}}\right)$. We explored τ in the range of $(0, 5]$ and $log_{10}(\nu)$ in the range of $[-2, 2]$.

For the 3-epoch model, there were 4 parameters; $\tau_1, \tau_2, \nu_2, \nu_3$. The subscripts refer to the epoch being referenced, with epoch 1 being the current epoch and epoch 3 being the oldest epoch. In this context, τ_1 is the duration of the first epoch (or the amount of time elapsed since epoch 2) and τ_2 is the duration of the second epoch (or the amount of time elapsed between epochs 1 and 3). The parameters ν_2, ν_3 refer to $\frac{N_{epoch\ 2}}{N_{epoch\ 1}}$ and $\frac{N_{epoch\ 3}}{N_{epoch\ 1}}$, respectively. The parameters for the 3-epoch model were explored for the same ranges as those for the 2-epoch model.

Neither scenario provided sufficient evidence to reject the null hypothesis of a standard neutral model at equilibrium. The MLE for the 2-epoch model fails to reject with P = 0.39 while the MLE of the 3-epoch model fails to reject at a P = 0.07. As a result, we used the standard neutral model without demography as the null hypothesis in tests of selection below.

7.2 Estimating the effects of ascertainment bias

The SFS may also be influenced by ascertainment bias. Because tiling array probes are designed by Affymetrix to avoid redundancy in the reference genome, they are all single-copy. In fact, in updating the annotation of the probes, we discarded the few remaining probes on the chips

that the release 4 genome assembly indicate are redundant. As a result, we were forced by the platform to limit our inferences on copy number changes segregating in non-redundant portions of the reference genome. Similarly, because deletions segregating in the reference genome are not present on the chip, we cannot assay those deletions. As a result, using microarrays creates a strong ascertainment bias. In order to correct for this, we propose the following model of ascertainment bias.

Given $m = n + 1$ individuals (n experimental lines and 1 reference line), no mutations falling in the reference line can be ascertained. As a result, the influences of ascertainment bias on the the i^{th} frequency class of the SFS can be modeled as follows:

$$E[y_i] = S\frac{E[x_i]\frac{m-i}{m}}{\sum_{i=1}^{m-2} E[x_i]\frac{m-i}{m}}, \quad i \leq m-2 \tag{3}$$

$E[x_i]$ is the expectation for the i^{th} frequency class of the SFS under a particular model, in this case as described (*S30*) above. $E[y_i]$ is the expectation for the i^{th} frequency class of the SFS after incorporating ascertainment bias. The numerator is proportional to the probability of a mutation being absent in the reference line. Thus, high frequency variants are less likely to be absent in the reference line than low frequency variants. As a result, the ascertainment biases cause a systematic under-sampling of high frequency variants. The S term is the total number of observed sites. The denominator and S terms are normalization factors ensuring that:

$$\sum_{i=1}^{n-1} E[y_i] = \sum_{i=1}^{n-1} E[x_i] = S \tag{4}$$

7.3 Estimating the effects of error

Because false positives and false negatives also influence the SFS we observe, we also modeled their effects. First, we modeled the influence of false negatives. Under a model of error, the

observed i^{th} frequency class of the SFS was composed of the true i^{th} frequency class mutations not affected by the false negative rate and all of the mutations in higher frequency bins effected by the false negative rate. For example, in a sample of 10 individuals, the observed 7^{th} frequency bin would be composed of the sites from the 7^{th} mutation bin not influenced by false negatives plus the number of 8^{th} frequency class sites where exactly one was effected by false negatives plus the number of 9^{th} frequency class sites where exactly two lines were effected by false negatives. If we assume that the error rate was independent, this can be modeled as a sum of binomials:

$$E_{TP}\left[z_i\right] = E\left[y_j\right] B\left(0, p_{FN}\right), \;\; 1 \leq i \leq n-1 \tag{5}$$

$$E_{FN}\left[z_i\right] = \sum_{j=i+1}^{n-1} E\left[y_j\right] B\left(j-i, p_{FN}\right), \;\; 1 \leq i \leq n-1 \tag{6}$$

Where $E_{TP}\left[z_i\right]$ is the expected number of true positives observed after error in the i^{th} frequency class, $E_{FN}\left[z_i\right]$ is the expected number of false negatives observed after error in the i^{th} frequency class, $B\left(k, p\right)$ is the binomial distribution, and p_{FN} is the false negative rate.

When calculating our empirical false positive rate, we used the following formulation:

$$p_{FP} = \frac{S_{FP}}{S_{Observed}} = \frac{S_{FP}}{S_{TP} + S_{FP}} \tag{7}$$

Where FP represents false positives, and TP represents true positives. Rearranging, we obtain the expected number of false positives:

$$E_{FP} = S_{FP} = S_{TP}\frac{p_{FP}}{1-p_{FP}} = \sum_{i=1}^{n-1} E\left[x_i\right] \frac{p_{FP}}{1-p_{FP}} = S\frac{p_{FP}}{1-p_{FP}} \tag{8}$$

Finally, we assume that false positives are added only to the singleton category ($i = 1$). As

a result, our final correction for error becomes:

$$E\left[z_i\right] = \begin{cases} S\frac{E_{FP}+E_{TP}[z_i]+E_{FN}[z_i]}{E_{FP}+\sum_{i=1}^{n-1}(E_{TP}[z_i]+E_{FN}[z_i])}, & i=1 \\ \\ S\frac{E_{TP}[z_i]+E_{FN}[z_i]}{E_{FP}+\sum_{i=1}^{n-1}(E_{TP}[z_i]+E_{FN}[z_i])}, & i>1 \end{cases} \tag{9}$$

We make two simplifying assumptions above, namely that false negatives are independent and that false positives contribute exclusively to the singleton category. The first assumption causes a strong reduction in the frequency for large i compared to smaller i while the second results in an increased singleton caterory. Since the consequences these assumptions mimic the effects of purifying selection on the SFS, they make purifying selection more difficult to infer. Thus, it is conservative to make such assumptions. Relaxing the assumptions, in addition to being much more complicated, would result in a less conservative estimate of the influence of error on the SFS. One essential caveat of this approach is that it is not conservative with respect to positive selection. If the ascertainment bias and bias introduced by error is weaker than we assumed, then we biased our results towards detecting positive selection. As a result, we urge caution to any who attempt to ascertain positive selection with the shape of the SFS in the presence of the types of biases we describe here.

Finally, we note that the influence of error on the expected SFS becomes overpowering for high values of p_{FP} and p_{FN}, especially when the number of sites S sampled is low. One consequence of this problem is that we could not use the shape of the SFS to examime natural selection on deletions under the models we propose here, as the error rates and sample sizes for deletions are so small as to render any PRF-SFS inferences meaningless. In fact, when we subject the deletions to PRF-SFS analysis, their confidence intervals are compatible with an extremely large range of parameter values for γ, both negative and positive. Moreover, the same problem would be encountered if existing categories were subdivided too finely. With this in mind, we have presented our data as finely subdivided as possible while still allowing us to conduct meaningfully powerful tests.

7.4 Testing hypotheses of natural selection

To test for the presence of natural selection, for each partition of the data (chromosome$\times$mutation type and annotation type$\times$mutation type), we set the demographic parameters to reflect a 1-epoch model as justified above and iterated over a parameter grid for the scaled selection coefficient γ using the multinomial model as described in Williamson *et al.* (*S30*) in order to obtain $E\left[x_i\right]$ for a sample size of 11, which is meant to represent the 10 natural lines we surveyed plus the reference genome. Typically, this expectation would be used to calculate the $\ln(L)$ of data of the data in the following way:

$$\ln\left(L\right) = \sum_{i=1}^{n-1} O\left[x_i\right] \ln\left(\frac{E\left[x_i\right]}{\sum_{i=1}^{n-1} E\left[x_i\right]}\right) \tag{10}$$

Where $O\left[x_i\right]$ is the observed number of sites in the i^{th} frequency class and $E\left[x_i\right]$ is the expected number under a PRF model. However, this formulation neglects the effects of ascertainment and error. Adjusting the expectations of the SFS based on a particular model of ascertainment bias and error can incorporate these forces directly into the likelihood ratio test inferences. Therefore, instead of the formulation in Equation 10, we first incorporate ascertainment bias and error into the expectation of the SFS. Subjecting the expected SFS to correction for ascertainment bias reduced the expectation from a sample size of 11 to a sample size of 10, which represents discarding the reference line, as no mutations can be ascertained there. The expection under a PRF model and an ascertainment bias model was then expressed as $E\left[y_i\right]$. After correcting for error, we obtained the expectation of the SFS in terms of $E\left[z_i\right]$, which simply replaces $E\left[x_i\right]$ in Equation 10 above.

From these modified expectations, we obtained both the maximum likelihood estimate (MLE) for γ as well as the parameter range for the 95% confidence interval. The 95% confidence interval was obtained within a likelihood ratio testing (LRT) framework. We determined what $\ln\left(Likelihood\right)$ difference would lead to a rejection of the null hypothesis at an error rate

of $\alpha = 0.05$ by examining the χ^2 distribution:

$$X^2 = 2 \times \Delta \ln (Likelihood) \tag{11a}$$

$$\chi^2_{0.95,df=1} = 3.84 \tag{11b}$$

$$\Delta \ln (Likelihood)_{Critical} = \frac{\chi^2_{0.95,df=1}}{2} = 1.92 \tag{11c}$$

Thus, the 95% confidence intervals reported reflect the range of parameter values within 1.92 likelihood units of the MLE.

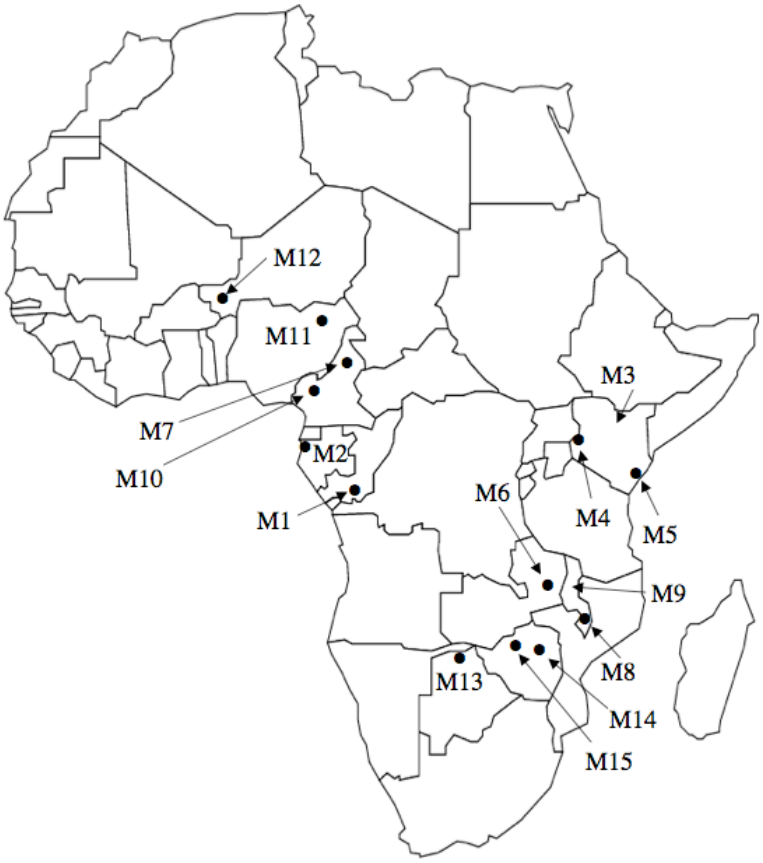

Figure S1: **Geographic distribution of natural *D. melanogaster* populations used in this study.** The *melanogaster* lines were derived from sub-Saharan Africa and are indicated by numbers whose names are listed in Table S1.

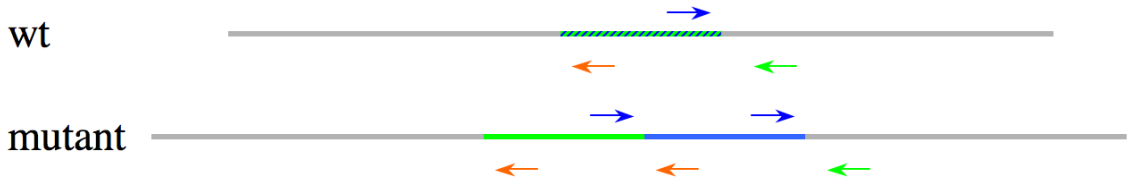

Figure S2: **Confirmation strategy for duplications.** In order to confirm duplications identified by the HMM, we designed primers to regions within the putative duplications. Each line represents the genome in a region of putative tandem duplication. The unduplicated regions are represented by gray segments. In the wild-type genome (top), the region that is hatched green and blue represents the region duplicated in the mutant genome (bottom). In the mutant genome, the $5'$ most copy is in green and the $3'$ most is represented in blue. The diagnostic primers are designed as indicated by the orange and blue arrows as indicated in the figure. With respect to the single copy genome (top) the primers are divergent and consequently cannot initiate a PCR reaction. However, when a tandem duplication occurs (bottom) the orange and blue primers become convergent, allowing amplification. Notably, even if the duplication is accompanied by inversion, primers of the same color will come into convergence, allowing amplification (not shown).

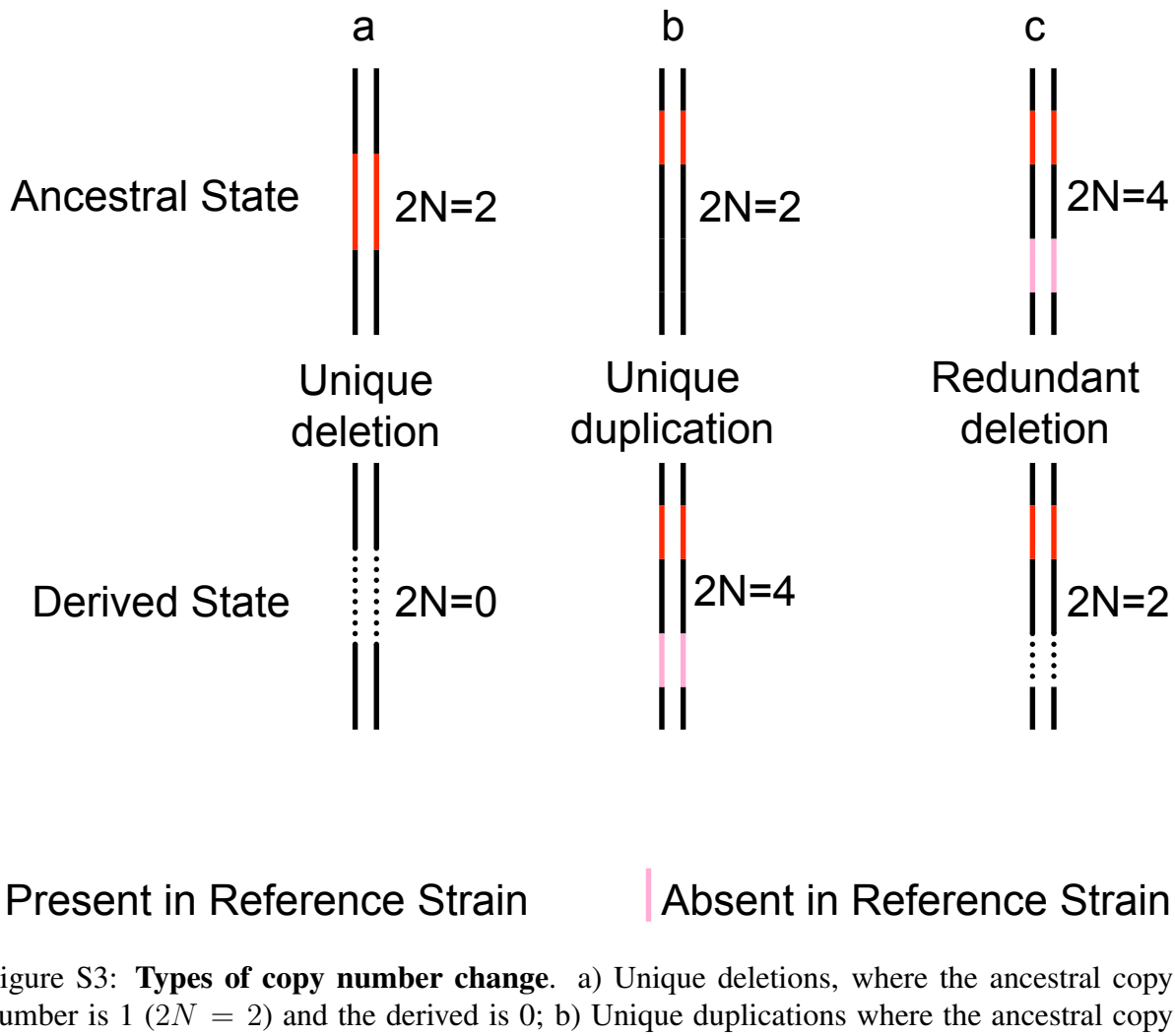

Figure S3: **Types of copy number change**. a) Unique deletions, where the ancestral copy number is 1 ($2N = 2$) and the derived is 0; b) Unique duplications where the ancestral copy number is 1 ($2N = 2$) and the derived is 2 fold higher ($2N = 4$); c) Redundant deletions where the ancestral copy number is 2 ($2N = 4$) and the derived copy number is half that ($2N = 2$). Black segments refer to regions of the genome that don't change. Red segments denote regions of the genome that are present in single copy in the reference genome but are polymorphic in our natural lines for either more ($2N = 4$) or fewer ($2N = 0$) copies. The pink segments are paralogous copies of the red segments and are absent in the reference genome. The dotted segments denote regions of the genome lost by a deletion event.

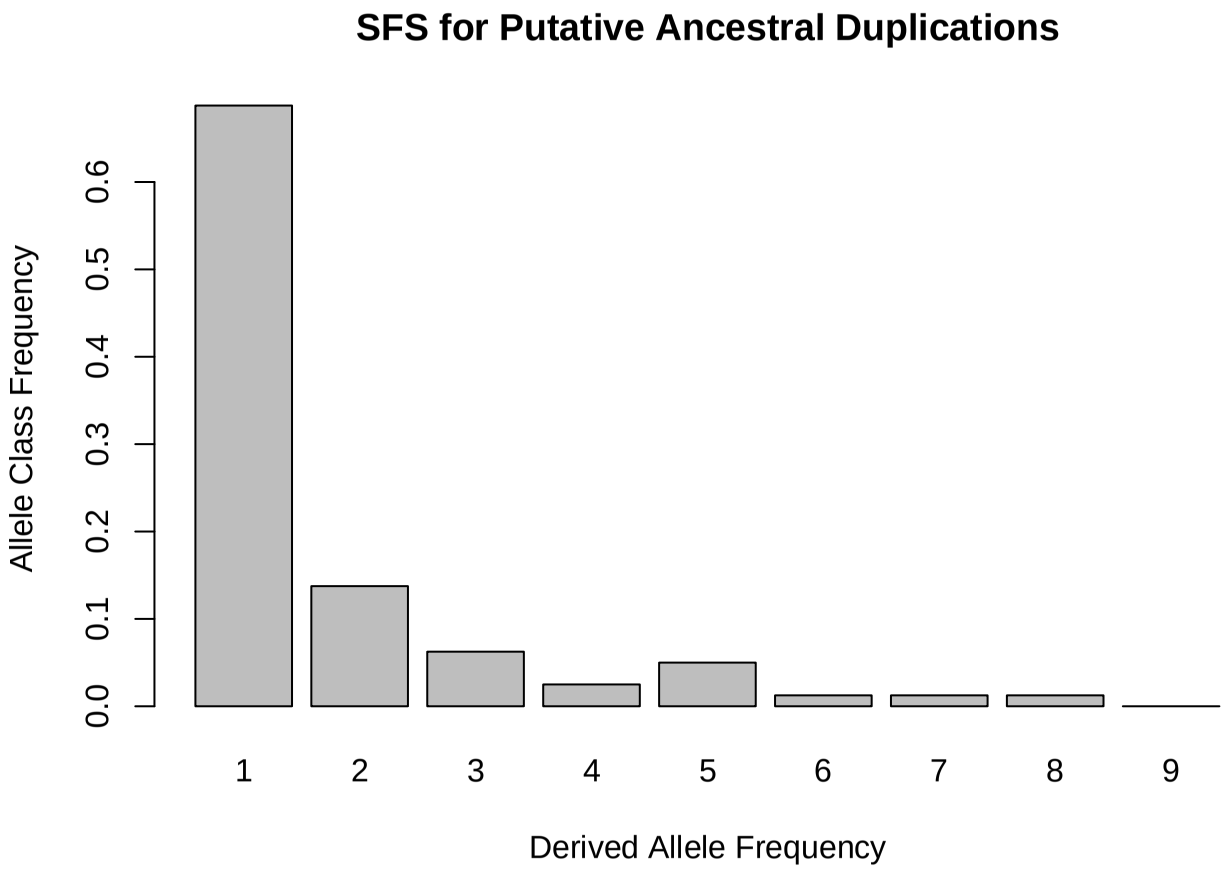

Figure S4: **Site frequency spectrum for putatively ancestral duplications**. Histogram showing the distribution of polymorphic duplications for all duplications showing evidence of ancestral redundancy in all chromosomes. The polarization of the mutations is plotted assuming that the two copy state is derived and the one copy state is ancestral.

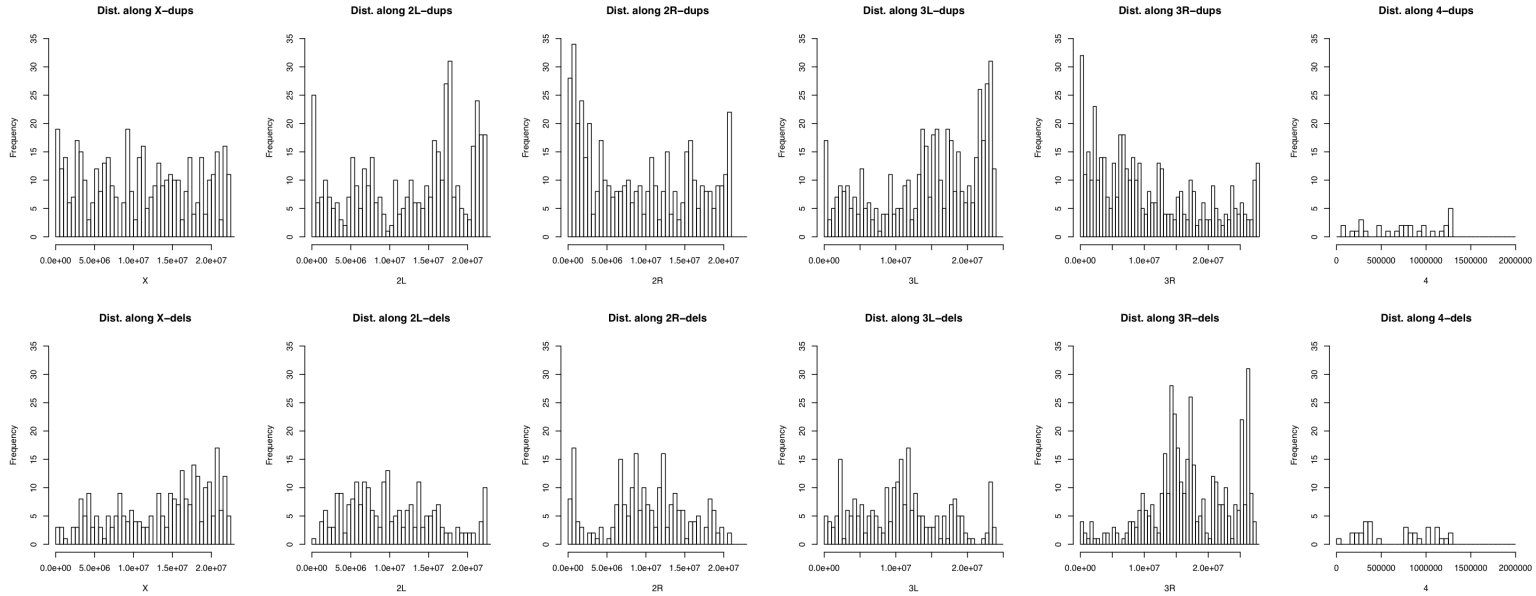

Figure S5: **Chromosomal distribution of CNPs**. Histogram showing the distribution of polymorphic duplications (top) and polymorphic deletions (bottom) for all chromosomes. Chromosomes 2 and 3 are divided into chromosome arms. The centromere is located between the left and right arms of chromosome 2 (2L & 2R) and chromosome 3 (3L & 3R). Each bin corresponds to 500kb of sequence for all chromosomes except for chromosome 4 where each bin corresponds to 50kb. Pericentromeric regions correspond to roughly 5 bins.

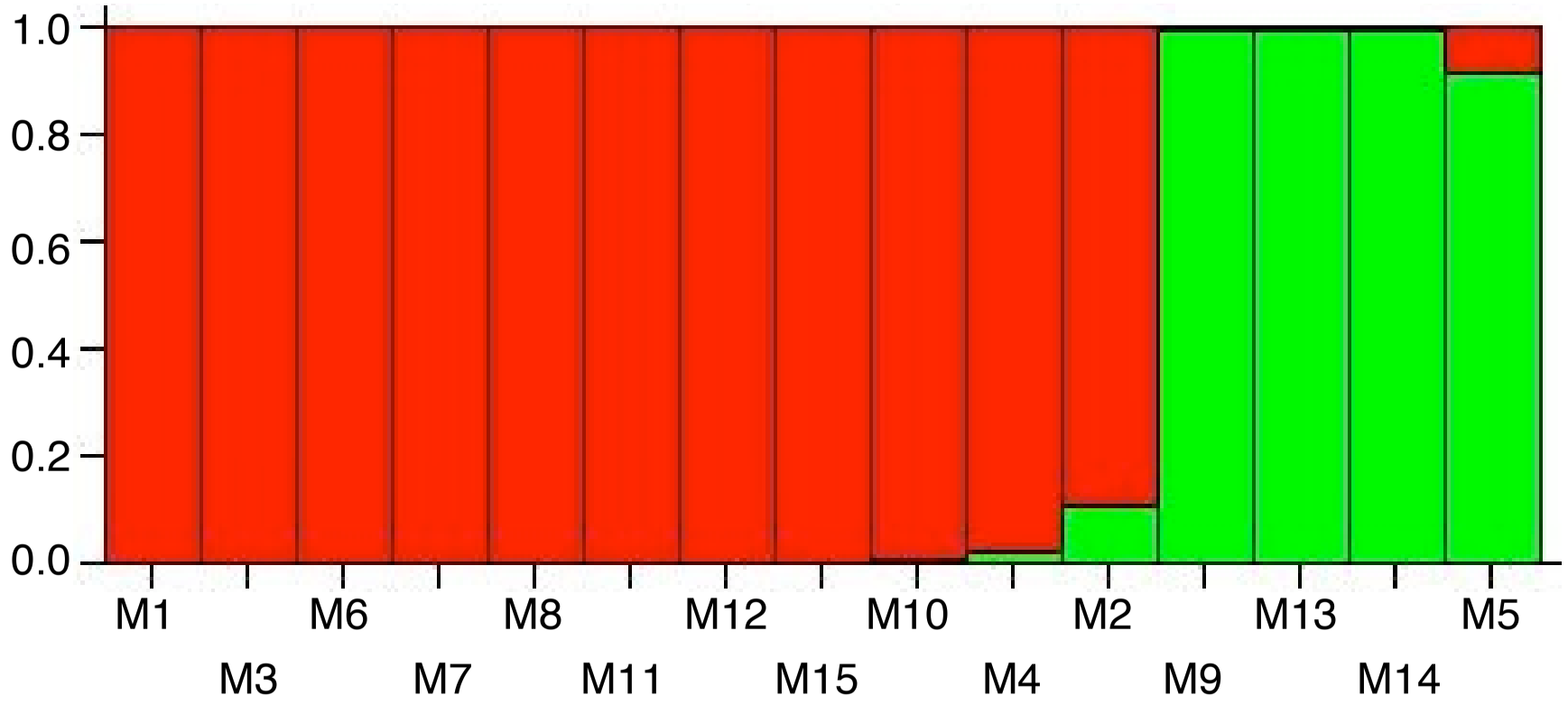

Figure S6: **Summary of population structure for silent SNP in *D. melanogaster*.** The figure illustrates the results for running structure on our silent SNP dataset. Notably, the first 10 individuals come almost universally from one population while the 12^{th}-14^{th} individuals come from another. The 11th and 15th individuals are primarily from one population or the other but appear to have significant, albeit low, levels of admixture.

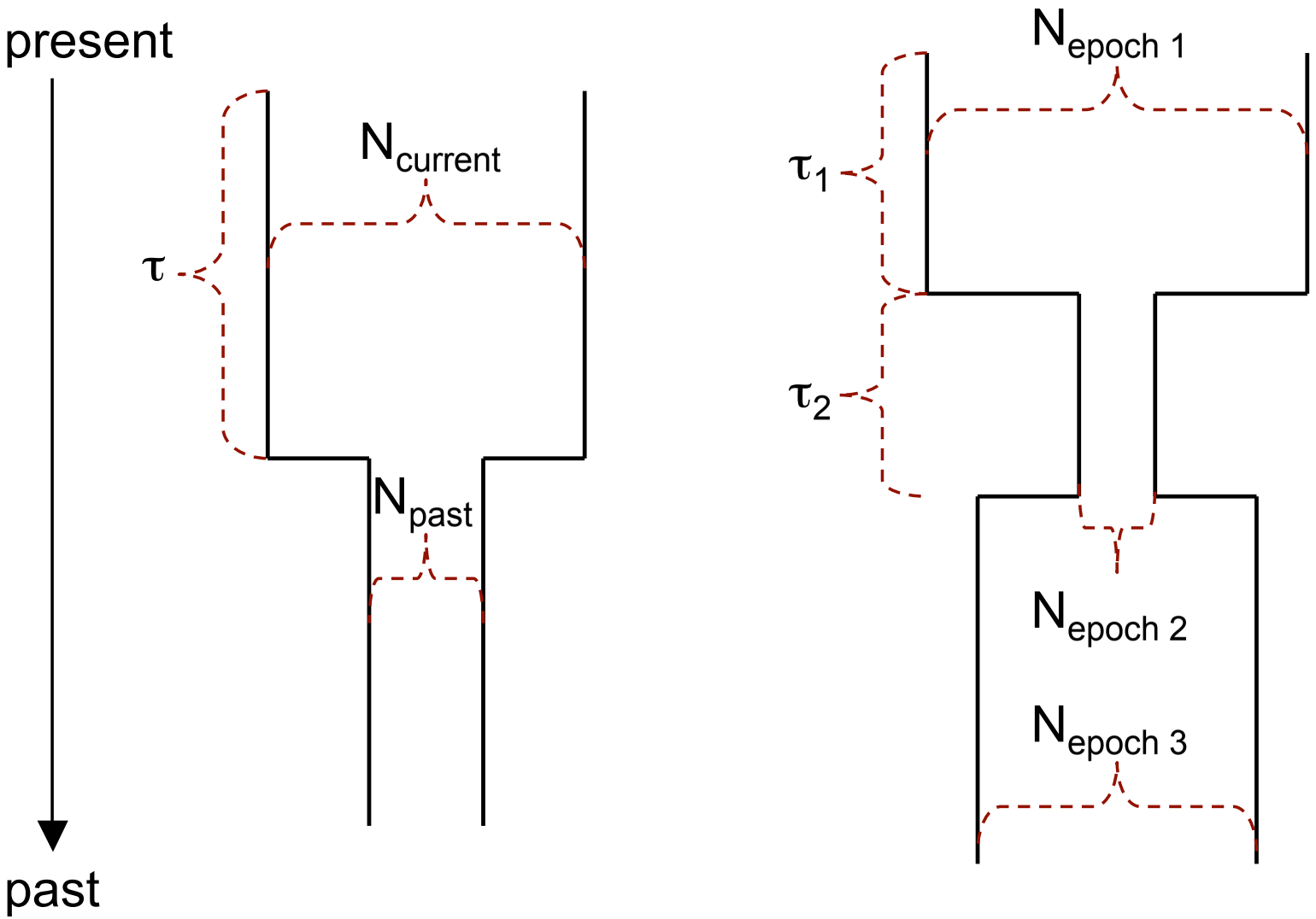

Figure S7: **Summary of demographic models explored.** The diagram illustrates graphically the relationships between the parameters that describe either a 2-epoch (left) or a 3-epoch (right) demographic model. The population sizes during various epochs are represented by the width between the lines (labeled with parameters involving N) while the time parameters are represented by the lenghts of the segments labeled by τ.

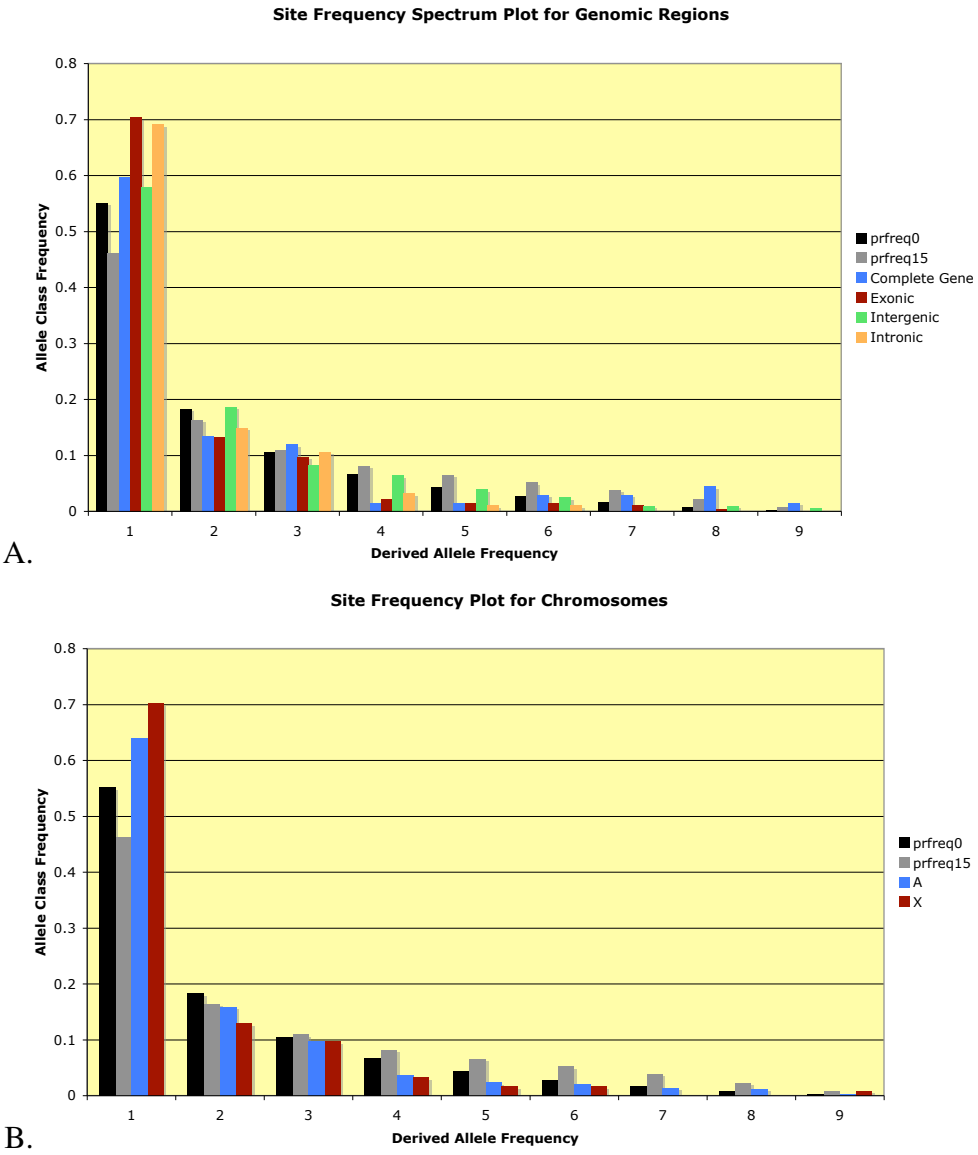

Figure S8: **Site frequency spectra for duplications in the *D. melanogaster*genome.** The figures show the site frequency spectrum for duplications of various partitions. In each subfigure, the black bar represents our most conservative null hypothesis that takes into account demography, ascertainment bias, and error (both false positives and false negatives). A) Comparison of the SFS between different annotation partions in the genome. The partitions mirror those in Fig. 1 in the main text. B) Comparison of the SFS between different chromosomes in the genome. "prfreq0" is the expected SFS under bias and error for $\gamma = 0$, while "prfreq15" is the same for $\gamma = 15$.

Name	Location	Map	Source	Major Pop.	Minor Pop.
Co	Congo (Brazzaville)	M1	1	1	-
Ga	Gabon (N'Toum)	M2	1	1	2
Ken48	Kenya	M3	2	1	-
KK	Kenya (Kakamega)	M4	1	1	-
KM	Kenya (Malindi)	M5	1	2	1
La79	Zambia (Luangwa)	M6	2	1	-
MD	Cameroon (Mbalang-Djalingo)	M7	1	1	-
ML	Malawi (Mwanza)	M8	1	1	-
MW6	Malawi	M9	2	2	-
NK	Cameroon (Nkouondja)	M10	1	1	-
NM	Nigeria (Maiduguri)	M11	1	1	-
Nr	Niger (Kareygorou)	M12	1	1	-
OK17	Botswana (Okavango Delta)	M13	2	2	-
ZH1	Zimbabwe (Harare)	M14	2	2	-
ZS5	Zimbabwe (Sengwa)	M15	2	1	-

Table S1: **Lines used in this study.** Name: the names of the natural *D. melanogaster* lines used in this study; Location and Map: their geographical locations with respect to the map on Fig. S1; Source: the provider of the lines; Major Pop.: the sub-population from which most of the SNP loci are derived; Minor Pop.: the sub-population from which the minority of SNP loci are derived, if any. For sources: 1 indicates lines obtained from collections made by available by John Pool (*S3*); 2 indicates lines obtained from the stocks in the laboratory of Chung-I Wu (*S4*), with lines M14 and M15 first used by the Aquadro lab in (*S5*). The populations in "Major Pop." and "Minor Pop." refer sub-populations inferred from STRUCTURE.

K	$\ln[P(D)]$	$var\{\ln[P(D)]\}$	$\alpha \pm 1$	$F_{st,1}$	$F_{st,2}$	$F_{st,3}$
1	-7930.6	302.1	-	0.0018	-	-
1	-7939.7	322.9	-	0.0004	-	-
1	-7943.3	327.2	-	0.0009	-	-
2	-7883.4	1446.2	0.0537	0.0008	0.3938	-
2	-7754.2	1250.6	0.0567	0.0282	0.3861	-
2	-7865.9	1410.1	0.0519	0.0022	0.3934	-
3	-9886.3	6065.8	0.1279	0.2497	0.427	0.2429
3	-9787.1	5877.3	0.1293	0.427	0.2407	0.2513
3	-9730.4	5774.6	0.1249	0.4249	0.2425	0.2508

Table S2: **Population structure; Representative results from different runs of STRUCTURE.** We include 3 runs of STRUCTURE each for numbers of subpopulations (K) ranging from 1 to 3. For each run, the burn-in length was 100,000 iterations of the Markov chain and the run time was an additional 100,000 iterations.

Table S3: **CNP calls for the *D. melanogaster* genome.** The spreadsheet describes the coordinates of mutations called by the criteria described above in Section 5.2. The start and end coordinate describe the first and last probes above the higher threshold. The context column describes what types of gene annotation the mutations overlap, following the designations in Fig. 1 in the main text.

Table S4: **True breakpoints obtained for CNPs following sequencing.** The spreadsheet compares the coordinates estimated from the model and those obtained empirically following sequencing.

Table S5: **List of genes completely duplicated or deleted.** The spreadsheet lists all genes duplicated/deleted in its entirety. For each gene there is information regarding if there are other genes completely mutated in the same event, chromosome location, known biological function and presence of paralogs in the genome.

Table S6: **List of the genes used to collect silent SNPs and their corresponding genomic locations.** The spreadsheet lists all silent SNP locations. The last column references the original article where these loci where first surveyed. Reference S24 reports that exon size influences the effectiveness of selection in that selection is less effective in long exons when compared to small exons as a result of Hill-Robertson interference. Hence, we only collected polymorphism data from those loci that showed the least amount of constraint (i.e., long exons).

Table S7: **Site frequency spectra of duplications and deletions.** The spreadsheet shows the distribution of duplications and deletions among the 15 lines used in this study partitioned by chromosome arm and genomic context.

Supplemental References and Notes

S1. M. D. Adams, *et al.*, *Science* **287**, 2185 (2000).

S2. G. Grumbling, V. Strelets, *Nucleic Acids Res* **34**, D484 (2006).

S3. J. E. Pool, C. F. Aquadro, *Genetics* **174**, 915 (2006).

S4. H. Hollocher, C. T. Ting, F. Pollack, C. I. Wu, *Evolution* **51**, 1175 (1997).

S5. D. J. Begun, C. F. Aquadro, *Nature* **365**, 548 (1993).

S6. M. Noe, Establishing a protocol for the detection of gene duplications in *Drosophila melanogaster*, BA honors thesis, Kalamazoo College (2004).

S7. J. J. Emerson, Evolution of genomic novelties, Ph.D. thesis, University of Chicago (2006).

S8. F. Biemar, *et al.*, *Proc Natl Acad Sci U S A* **102**, 15907 (2005).

S9. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, *J Comput Biol* **7**, 203 (2000).

S10. J. O. Borevitz, *et al.*, *Genome Res* **13**, 513 (2003).

S11. L. Rabiner, *Proceedings of the IEEE* **77**, 257 (1989).

S12. R. Durbin, S. R. Eddy, K. Anders, M. Graeme, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998).

S13. We adopted this convention, not followed in references cited here, in order to disambiguate the indices i and j, such that the focal position in the recurrence equation is always indexed by j. Other conventions force one to change the focal index to i in some contexts. Although the amount of notation increases slightly with the addition of the index k, the

improvement in clarity and the added ability to directly compare equations easily offsets this small cost.

S14. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics* **19**, 185 (2003).

S15. V. G. Tusher, R. Tibshirani, G. Chu, *Proc Natl Acad Sci U S A* **98**, 5116 (2001).

S16. A. Dempster, N. Laird, D. Rubin, *J. Roy. Stat. Soc.* **39**, 1 (1977).

S17. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, *Proc Natl Acad Sci U S A* **100**, 11484 (2003).

S18. S. Aulard, J. R. David, F. Lemeunier, *Genet Res* **79**, 49 (2002).

S19. E. Baudry, B. Viginier, M. Veuille, *Mol Biol Evol* **21**, 1482 (2004).

S20. P. R. Haddrill, K. R. Thornton, B. Charlesworth, P. Andolfatto, *Genome Res* **15**, 790 (2005).

S21. H. Li, W. Stephan, *PLoS Genet* **2**, e166 (2006).

S22. F. Tajima, *Genetics* **123**, 585 (1989).

S23. P. Andolfatto, *Nature* **437**, 1149 (2005).

S24. J. M. Comeron, T. B. Guthrie, *Mol Biol Evol* **22**, 2519 (2005).

S25. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* **8**, 175 (1998).

S26. B. Ewing, P. Green, *Genome Res* **8**, 186 (1998).

S27. D. Gordon, C. Abajian, P. Green, *Genome Res* **8**, 195 (1998).

S28. D. A. Nickerson, V. O. Tobe, S. L. Taylor, *Nucleic Acids Res* **25**, 2745 (1997).

S29. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* **155**, 945 (2000).

S30. S. H. Williamson, *et al.*, *Proc Natl Acad Sci U S A* **102**, 7882 (2005).