

**Elevation of linkage disequilibrium above neutral expectations in ancestral and derived populations of *Drosophila melanogaster***

Nandita R. Garud<sup>\*†</sup> and Dmitri A. Petrov<sup>‡</sup>

\* Department of Genetics, Stanford University, Stanford, CA 94305, USA

† The Gladstone Institute at the University of California, San Francisco, CA 94158, USA

‡ Department of Biology, Stanford University, Stanford, CA 94305, USA

Running title: Elevation of LD in *D. melanogaster*

KEYWORDS: *D. melanogaster*, demography, linkage disequilibrium, haplotype homozygosity, selection

Corresponding Authors:

Nandita R. Garud, PhD  
The Gladstone Institutes  
1650 Owens Street  
San Francisco, CA 94158  
(415) 734-2000  
Nandita.Garud@gmail.com

Dmitri A. Petrov, PhD  
Department of Biology  
Stanford University  
Stanford, CA 94305  
(650)-736-1169  
dpetrov@stanford.edu

## ABSTRACT

The extent to which selection and demography impact patterns of genetic diversity in natural populations of *Drosophila melanogaster* is yet to be fully understood. We previously observed that linkage disequilibrium (LD) at scales of ~10 kb in the Drosophila Genetic Reference Panel (DGRP), consisting of 145 inbred strains from Raleigh, North Carolina, measured both between pairs of sites and as haplotype homozygosity, is elevated above neutral demographic expectations. We also demonstrated that signatures of strong and recent soft sweeps are abundant. However, the extent to which these patterns are specific to this derived and admixed population is unknown. It is also unclear whether these patterns are a consequence of the extensive inbreeding performed to generate the DGRP data. Here we analyze LD statistics in a sample of more than 100 fully sequenced strains from Zambia; an ancestral population to the Raleigh population that has experienced little to no admixture and was generated by sequencing haploid embryos rather than inbred strains. This data set allows us to determine whether patterns of elevated LD and signatures of abundant soft sweeps are generic to multiple populations of *D. melanogaster* or whether they are generated either by inbreeding, bottlenecks or admixture in the DGRP dataset. We find an elevation in long-range LD and haplotype homozygosity compared to neutral expectations in the Zambian sample, thus showing the elevation in LD is not specific to the DGRP data set. This elevation in LD and haplotype structure remains even after controlling for possible confounders including genomic inversions, admixture, population substructure, close relatedness of individual strains, and recombination rate variation. Furthermore, signatures of partial soft sweeps similar to those found in the DGRP as well as partial hard sweeps are common in Zambia. These results suggest that while the selective forces and sources of adaptive mutations may differ in Zambia and Raleigh, elevated long-range LD and signatures of soft sweeps are generic in *D. melanogaster*.

## INTRODUCTION

Disentangling the effects of demography and selection on patterns of genomic variation remains a central challenge in evolutionary biology. Until recently, inference of demography and selection relied primarily on short genomic fragments sampled from a limited number of individuals (e.g., DUCHEN *et al.* 2013; GUTENKUNST *et al.* 2009; LI AND STEPHAN 2006; MOLINA *et al.* 2001; PRITCHARD *et al.* 2000; THORNTON AND ANDOLFATTO 2006). While these studies provided important insights into the evolutionary forces acting on populations, they were ultimately limited by both sample size and the physical scale in which patterns of polymorphism and linkage could be investigated. The recent availability of whole-genome sequences from multiple individuals with populations in a variety of species (e.g., ABECASIS *et al.* 2010; CAO *et al.* 2011; MACKAY *et al.* 2012) is enabling us to examine long-range patterns of linkage disequilibrium (LD) (~10Kb), measured either as correlations between pairs of sites or as haplotype homozygosity. LD offers powerful insights into selective and demographic processes shaping genetic variation in a natural population, such as bottleneck or admixture events (PENNINGS AND HERMISSON 2006; POKALYUK 2012; WALL AND PRITCHARD 2003; ANDOLFATTO AND PRZEWORSKI 2000). In addition, haplotype homozygosity is very sensitive to recent and strong adaptation leading to both classical hard and soft selective sweeps (GARUD *et al.* 2015; GARUD AND ROSENBERG 2015; PENNINGS AND HERMISSON 2006).

In the case of *Drosophila melanogaster*, the most recent demographic models were inferred using short intergenic and intronic sequence fragments on the X chromosome of lengths no longer than a few hundred base pairs in tens of individuals (DUCHEN *et al.* 2013). Because the data were so sparse, models assuming independence between polymorphic sites were fit to statistics sensitive to the site frequency spectrum and to short-scale LD measuring the correlation between polymorphic sites separated by short distances (DUCHEN *et al.* 2013).

The availability of the *Drosophila* Genetic Reference Panel (DGRP) (version 1), consisting of 168 fully sequenced *D. melanogaster* strains from a single population in Raleigh, North Carolina (MACKAY *et al.* 2012), offered the opportunity to study genome-wide signatures of demography and selection in a deep sample. Recently, we analyzed patterns of LD and haplotype homozygosity at distances of ~10Kb in the DGRP data set (GARUD *et al.* 2015) and compared these patterns with neutral expectations generated under several demographic models

of North American *D. melanogaster*, including two constant  $N_e$  models, two bottleneck models, and two recently inferred admixture models (DUCHEN *et al.* 2013), one of which is considered to be the most accurate model of the admixture between European and African flies founding North American *D. melanogaster*. We found that LD and haplotype homozygosity in the DGRP data is elevated relative to expectations under any of the demographic models tested and that this elevation remained even after controlling for several potential sources of elevated LD, including close relatedness between individuals, presence of genomic inversions, recombination rate variation, and population substructure.

In addition, we identified regions in the DGRP data with the elevated haplotype homozygosity, measured using statistic H12, designed to detect both hard and soft sweeps GARUD *et al.* 2015. H12 calculates haplotype homozygosity after combining the frequencies of the first and second most common haplotypes in order to increase the power to detect soft sweeps. Among the three highest-ranking candidates in the H12 scan were the prominent known cases of soft sweeps at the loci *Ace*, *Cyp6g1*, and *CHKov1* (KARASOV *et al.* 2010; MAGWIRE *et al.* 2011; SCHMIDT *et al.* 2010). All 50 outlier H12 peaks in the DGRP data showed signatures of partial soft sweeps, suggesting that soft sweeps were common in the DGRP data (but see SCHRIDER *et al.* 2015).

The extent to which soft sweeps are specific to this Raleigh population remains unknown. Several factors challenge the interpretation of the results of the DGRP data set, and we sought to address them in this study. First, the Raleigh population, like all other North American populations of *D. melanogaster*, is a recently derived population resulting from admixture between European and African lineages of flies (BERGLAND *et al.* 2015; DUCHEN *et al.* 2013; POOL *et al.* 2012). The effects of admixture on patterns of LD and haplotype homozygosity are still yet to be fully understood, and furthermore, the true demographic history of North American flies is undoubtedly much more complex than our current models can possibly capture. For example, POOL 2015 showed that epistatic interactions between African alleles and European alleles are abundant in the North American population and that these combinations of alleles appear to be experiencing selection against them.

Second, the flies composing the DGRP data set were extensively inbred to result in mostly homozygous genomes. This inbreeding process left behind tracts of residual heterozygosity, which were suggested by HOULE AND MARQUEZ 2015 to result from epistatic

interactions between inversions (particularly heterozygous inversions that may be lethal as homozygotes) maintained on different chromosomes or other deleterious mutations in repulsion. Unfortunately, this may result in elevations in long-range linkage disequilibrium, since regions of the genome bearing deleterious mutations may have to be maintained in-sync with other potentially non-deleterious regions of the genome. Otherwise, the compound effect of two deleterious regions present in the genome could be lethal.

Recently, the *Drosophila* Genome Nexus (DGN) dataset was published (LACK *et al.* 2015), consisting of 205 DGRP strains from Raleigh (inclusive of the original 168 DGRP strains available in the initial release) as well as 197 strains from Zambia. The Zambian data set offers the opportunity to test whether signatures observed in the Raleigh population are generic to multiple populations of *D. melanogaster*. Importantly, the two features of extensive admixture in the Raleigh population and inbreeding of the flies to generate the data set are absent from the Zambian lines. First, as the likely ancestral *D. melanogaster* population, the Zambian population has experienced relatively little admixture (POOL *et al.* 2012). Second, since Zambian flies have a much higher level of genetic diversity (POOL *et al.* 2012) and deleterious load, the flies cannot be fully inbred to generate phased haplotypes for the sequenced data set. Instead, LANGLEY *et al.* 2011 used a crossing technique to generate and sequence haploid embryos from isofemale strains. The lack of admixture and lower levels of inbreeding in the Zambian lines set make a comparison of patterns of LD and haplotype homozygosity between the Zambian data and the Raleigh data particularly informative in understanding general patterns of soft selective sweeps in *D. melanogaster*.

In this paper, we compared patterns of linkage disequilibrium and haplotype homozygosity in flies sampled from Raleigh and Zambia. We found that LD and haplotype homozygosity in both Raleigh and Zambian lines were elevated relative to neutral expectations, suggesting that elevated LD and haplotype homozygosity is a generic feature of multiple *D. melanogaster* populations. Furthermore, our H12 scan of the Zambian lines revealed several peaks in genomic locations different from that of the scan in Raleigh lines, many of which showed signatures of soft sweeps with multiple haplotypes at high frequencies and several which showed signatures of hard sweeps with one main haplotype at high frequency. In the Zambian data set, we recovered peaks at two positive controls, *Ace* and *Cyp6g1*, but these were not the highest-ranking peaks as in the Raleigh data set, reflecting the lower frequencies at which the

adaptive mutations are found in ancestral, African populations than in derived populations. We do not recover a peak at *CHKov1*, which is a negative control in African populations since the adaptive mutation has been shown to be present at high frequencies only in non-African populations (AMINETZACH *et al.* 2005). While the patterns of elevated long-range LD and multiple haplotypes at high frequencies in both populations show that soft sweeps may be common in both populations, the differences in genomic locations of individual selective sweeps were consistent with different underlying selective pressures and evolutionary processes in the two populations.

## MATERIALS AND METHODS

### *Data*

In this paper we contrast levels of LD and haplotype homozygosity in the Raleigh and Zambian populations measured along the four autosomal arms. In our analysis, we masked regions of identity by descent (IBD), region between genomic inversion breakpoints, and tracts of residual heterozygosity resulting from the inbreeding of the Raleigh strains, which were all previously inferred by CORBETT-DETIG AND HARTL 2012; HUANG *et al.* 2014; LACK *et al.* 2015. For Zambian strains, we also masked previously inferred regions of admixture from Europe into Africa (LACK *et al.* 2015; POOL *et al.* 2012) (Text S1). Finally, we excluded regions of the genome with low recombination rates  $< 1$  centimorgan/megabase (cM/Mb) (using female meiosis rates from COMERON *et al.* 2012). The extensive masking of this data set resulted in large tracts of missing data across several individuals and introduced variation in the sample size at different sites in the genome. To account for this variation, we sub-sampled both the Raleigh and Zambian data to 100 strains, excluding the worst strains in terms of missing data (Text S1). We refer to this second data set as the stringently filtered data set.

Later in this paper we perform an H12 scan of the Zambian and Raleigh data sets along the four autosomal arms to identify individual selective events in each population. To maximize our coverage and power to identify selective sweeps, we applied less stringent filters to the data by masking tracts of residual heterozygosity and additional heterozygous sites as before, but not masking any inversions, regions of inferred admixture, or IBD tracts. Instead, we *a posteriori* tested for enrichment of selective sweeps in genomic regions with inversions present at high frequency in the data set. In addition, instead of masking IBD tracts, we excluded all individuals with genome-wide IBD levels with at least one other individual at  $> 20\%$  (Text S1). This resulted in 178 and 188 strains in the Raleigh and Zambian data sets, respectively. We further down-sampled the data sets to 145 strains excluding the strains with the highest amount of missing data (Text S1) to match the sample depth of our previous H12 scan (GARUD *et al.* 2015). Finally, we excluded regions of the genome with recombination rates  $< 0.5$  cM/Mb (using female meiosis rates from COMERON *et al.* 2012). We refer to this second data set as the less stringently filtered data set.



### ***Demographic inference***

To assess whether LD and haplotype homozygosity levels observed in the Zambian and Raleigh data are elevated relative to neutral expectations, we measured LD and haplotype homozygosity in neutral models inferred for Zambian and Raleigh data. We inferred demographic models for the Zambian data using the program DaDi (GUTENKUNST *et al.* 2009) and used previously inferred demographic models for the Raleigh data (DUCHEN *et al.* 2013; GARUD *et al.* 2015). DaDi calculates a log-likelihood of the fit of a model based on an observed site frequency spectrum (SFS). We estimated the SFS for presumably neutral SNPs in Zambian data using segregating sites in short introns (PARSCH *et al.* 2010) from the stringently filtered data set. We used every site in a short intron of length less than 86 bps, with 16 bps removed from the intron start and 6 bps removed from the intron end (LAWRIE *et al.* 2013). Unlike the stringently filtered data set we used to calculate estimates of LD from, we did not remove any strains before inferring the models, and instead projected the SFS down to 100 samples at each polymorphic site. This procedure resulted in 73,197 SNPs in the Zambian data out of a total of 738,024 bps.

The models we fit to the Zambian short intron data include a bottleneck model, a growth model, a bottlegrowth model (a three epoch model with a bottleneck followed by a growth), and a constant  $N_e$  model (Table S1). For the bottleneck and bottlegrowth models, we inferred four free parameters: the bottleneck size ( $N_B$ ), the bottleneck time ( $T_B$ ), final population size ( $N_F$ ), and the final population time ( $T_F$ ). For the growth model, we inferred two free parameters: final population size ( $N_F$ ) and duration of the growth ( $T$ ). Finally, we also inferred a constant  $N_e = 4,789,108$  ( $\approx 4.8 \times 10^6$ ) for the constant population size model using a Watterson's theta estimate of the Zambian data and assuming a mutation rate of  $10^{-9}$ /bp/generation. Similarly, we inferred a constant  $N_e = 2,603,678$  ( $\approx 2.6 \times 10^6$ ) fit to Watterson's theta estimate for the short intron region of the stringently filtered Raleigh data, similar to the constant  $N_e$  model inferred by GARUD *et al.* 2015.

### ***Simulations of neutral demographic scenarios and selection scenarios***

We simulated population samples under the neutral demographic models fit to Zambian and Raleigh data with the coalescent simulator MSMS (EWING AND HERMISSON 2010). We simulated samples of size 100 with a recombination rate of 1 cM/Mb to match the sample depth and the

minimum recombination rate observed in the stringently filtered data. We assumed a neutral mutation rate of  $10^{-9}$  mutations/bp/gen (KEIGHTLEY *et al.* 2009).

We performed simulations of hard and soft sweeps for the Bayes factor analysis presented later in this paper. We used MSMS to simulate sample sizes of 145, matching the sample depth of the less stringently filtered data. We simulated hard and soft sweeps arising from *de novo* mutations by specifying the population parameter  $\theta_A = 4N_e\mu_A$  at the adaptive site to be 0.01 and 10, respectively, matching the parameter estimates used in GARUD *et al.* 2015. The adaptive site was always placed in the center of the locus. We performed  $10^6$  hard and soft sweep simulations in constant  $N_e=4.8*10^6$  and  $2.6*10^6$  scenarios matching the expected constant  $N_e$  population sizes for the Zambian and Raleigh populations, respectively. We assumed a neutral mutation rate of  $10^{-9}$  events/bp/gen (KEIGHTLEY *et al.* 2009) and recombination rates of 0.5 and 5 cM/Mb, reflecting the range of recombination rates observed in the data. We also assumed co-dominance, whereby a homozygous individual bearing two copies of the advantageous allele has twice the fitness advantage of a heterozygote. The selection coefficient ( $s$ ), partial frequency of the adaptive allele after selection ceases ( $PF$ ), and time since selection ceased ( $T_E$ ) were drawn from uniform prior distributions as follows:  $s \sim U[0,1]$ ,  $PF \sim U[0,1]$ , and  $T_E \sim U[0,1000]$  generations.

### ***Linkage disequilibrium estimates***

We measured LD in the stringently filtered data and in neutral simulations (Figures 1, S1, and S2). LD was measured using the  $r^2$  statistic. If any SNP had missing data in the stringently filtered data, the individuals with the missing data were excluded from the LD calculation as in JAKOBSSON *et al.* 2008. At least four individuals without missing data at both SNPs were required to compute LD, otherwise the SNP pair was discarded.

We smoothed the LD plots in Figures 1, S1, and S2 by averaging LD values binned according to the distance between SNPs. We binned LD values initially in non-overlapping 20-bp windows until a distance of 300 bps. After that, LD values were binned in 150-bp non-overlapping windows. We chose two different bin sizes for short range and long range LD because LD decays rapidly at short distances.

### ***Genomic scan for selective sweeps in Zambian and Raleigh data using H12***

We scanned the autosomal arms of the Zambian and Raleigh genomes in the less stringently filtered data sets for selective sweeps using the haplotype homozygosity statistic H12, designed to detect hard and soft sweeps (GARUD *et al.* 2015) (Figure 2). For the Zambian data we used analysis windows of 801 SNPs randomly down sampled to 401 SNPs (Figure 2a). For the Raleigh data, we used windows of 401 SNPs. The centers of the windows were separated by 50 SNPs. Haplotypes were grouped together only if they matched at all sites. Any haplotype with missing data matching multiple haplotypes at all sites except the sites with missing data were randomly grouped with one other matching haplotype.

To identify selective events, we called H12 peaks in the data. We grouped together consecutive tracts of H12 analysis windows lying above the median H12 value in the data and chose the highest H12 value among all windows in the peak to represent the H12 value of the entire peak. We defined the edge coordinates of each peak as the smallest and largest coordinates of all SNPs comprising a peak across all windows. After identifying all peaks, we iterated through the peaks in reverse order starting with the highest H12 value and excluded all peaks within 500Kb from the center of each peak until all peaks had been accounted for. This avoided the potential problem of identifying multiple peaks in the same genomic region belonging to the same selective event. Finally, we excluded peaks with edge coordinates overlapping any region identified by COMERON *et al.* 2012 with  $\rho < 0.5$  cM/Mb.

To identify selective events occurring in both the Zambian and Raleigh populations, we found overlapping peaks, where an overlap was defined as the intersection of coordinate pairs for the peaks identified in the Zambian and Raleigh data sets. For more details on calculating the probability of an overlap of peaks by chance in the Zambian and Raleigh scans, see Text S1.

## RESULTS

### *Slow decay of linkage disequilibrium in the Zambian and Raleigh populations relative to neutral expectations*

We compared the decay in LD in the Zambian data with expectations under neutral demographic models fit to the data (Figure 1a). We inferred four simple neutral demographic models fit to short introns in the Zambian lines using the program DaDi (GUTENKUNST *et al.* 2009): a constant  $N_e$  demographic model, a bottleneck model, a growth model, and a model of a bottleneck

followed by expansion (hereafter, ‘bottlegrowth’).  $S$  and  $\pi$  for each model match empirical estimates from short-intron data in the Zambian population (Table 1). The log-likelihoods of each model and their inferred parameters are presented in Table 2. Based on the log likelihoods, the bottlegrowth model and bottleneck models have the best fits. However, since even the bottleneck model experiences a growth in population size, we decided to focus on the bottlegrowth model. In addition, we estimated the decay in LD in simulations of the constant  $N_e$  and growth models as points of comparison. Since low recombination can result in elevated LD, we excluded regions of the genome with recombination rates  $< 1$  cM/Mb (COMERON *et al.* 2012) and performed all neutral simulations with this recombination rate. This recombination rate should be conservative for our purposes as this was the lowest recombination rate in our data.

We observed a substantial elevation in LD in the data relative to neutral expectations under all demographic models (Figure 1a). We estimated LD for different minor allele frequency (MAF) classes of SNPs in which both SNPs in an LD pair had MAFs between 0.05 and 0.5 (Figure 1a), 0.3 and 0.5 (Figure S1a), and 0 and 0.05 (Figure S1b). Using a two-sided Wilcoxon rank sum test to test the significance of the deviation in LD from neutral expectations, the  $P$ -value was  $< 10^{-15}$  in all cases. The deviation in LD in the data from expectations is apparent at short distances less than 100 base pairs, consistent with observations of PRZEWORSKI *et al.* 2001 for multiple species of *Drosophila*, as well as for long distances, consistent with our previous results (GARUD *et al.* 2015).

In Figures 1b, S1c and S1d, we reiterated our previous result in GARUD *et al.* 2015 in the Raleigh data set and confirmed that there is an elevation in LD in the Raleigh population relative to the six neutral demographic models fit to the Raleigh data in terms of  $S$  and  $\pi$  (Table 1). The six neutral demographic models fit to Raleigh data include an admixture model for North American flies (DUCHEN *et al.* 2013), a variant of the admixture model which includes a bottleneck in one of the contributing populations (DUCHEN *et al.* 2013), two constant  $N_e$  models, one which was fit to Watterson’s theta estimate using short intron data (Methods), and two bottleneck models we previously fit to the DGRP data (GARUD *et al.* 2015). Unlike the LD in the Zambian data set, the deviation in LD from neutral expectations is even more exaggerated in the Raleigh data set (for all MAFs conditioned on and demographic models tested, the  $P$ -value is  $< 10^{-15}$  for a two-sided Wilcoxon rank sum test). These findings are consistent with previous findings showing that the elevation in LD is muted in an ancestral Zimbabwean population

relative to non-African populations (ANDOLFATTO AND WALL 2003). The elevation of LD in both Zambian and Raleigh data suggest that elevated LD is a generic feature of multiple *D. melanogaster* populations.

To confirm that the elevation in LD is not driven by any anomaly in the data, such as unaccounted related individuals, inversions, or strains with unusually high polymorphism or sequencing error rates, we subsampled the Zambian and Raleigh data sets 10 times to 50 strains to test whether any subset of individuals may be driving the elevation in LD. In all subsampled data sets, LD remains elevated in contrast to expected LD measured in all tested demographic models simulated with a sample size of 50 ( $P$ -value  $< 10^{-15}$  for all scenarios tested, two-sided Wilcoxon rank sum test) (Figure S2), suggesting that the elevation in LD is driven by population-wide phenomena.

Finally, although we excluded the intervals between all inversion breakpoints from our stringently filtered data (Text S1), we tested whether LD within 100Kb of the inversion breakpoints is elevated relative to the rest of the genome since inversions can impact LD at long ranges. We found that the mean difference in LD in regions within 100Kb versus beyond 100Kb of inversions have little biological difference for both the Raleigh and Zambian populations (Table S3).

### ***Elevation of haplotype homozygosity in the Zambian and Raleigh populations relative to neutral expectations***

We compared the distribution of haplotype homozygosity measured with H12 in the Zambian and Raleigh data sets with the distribution of H12 values measured in neutral demographic scenarios fit to the two populations (Figures 1c and 1d). H12 is designed to detect both hard and soft sweeps and is calculated by combining the frequencies of the two most abundant haplotypes in an analysis window (GARUD *et al.* 2015). For the Raleigh data, we applied the statistic in analysis windows of 401 SNPs (200 SNPs on each side of the central SNP of each window) with the centers of each window iterated by 50 SNPs. 401 SNPs, corresponds to ~10Kb in the Raleigh data (GARUD *et al.* 2015). In the Zambian data, however, 801 SNPs corresponds to approximately ~10Kb. Therefore, we applied the statistic in 801 SNP windows randomly down sampled once to 401 SNPs in the Zambian data to have comparable analysis windows sizes in terms of SNPs and base pairs for both scans. As we show in Figures 1c and d, a window of this

length scale is unlikely to generate high values of LD or haplotype by chance under neutrality. We measured H12 in 150,000 simulations of each neutral demographic scenario, representing more than 10 times the number of analysis windows observed in the data.

Figures 1c and 1d show that there is a long tail of high H12 values in both the Zambian and Raleigh data sets relative to the distribution of H12 measured in simulations under all neutral demographic models tested, confirming our previous results (GARUD *et al.* 2015) that showed that all the neutral demographic scenarios we tested are unable to generate the haplotype structure observed in the Raleigh data by chance ( $P$ -value  $< 10^{-15}$ , Wilcoxon rank sum test, for all neutral demographic scenarios tested). The elevation in H12 values in the tail was muted in the Zambian data set with the highest H12 value reaching only 0.0936, whereas in the Raleigh data set, the highest value reached 0.5374. In the Raleigh data set, the median H12 value (0.0202) was sharply elevated above the median H12 value measured in simulations ( $\sim 0.012$  across all demographic scenarios). In the Zambian data set, we found that the median H12 value in the data (0.0124) is also elevated relative to the median H12 value measured in simulations ( $\sim 0.0109$  across all demographic scenarios). To identify analysis windows with H12 values significantly higher than expected under neutrality, we calculated critical values ( $H12_o$ ) under the constant  $N_e$  models fit to Zambian and Raleigh data using a 1-per-genome false discovery rate criterion (Text S1). We found that  $\sim 64\%$  and  $\sim 19\%$  of the windows in the Raleigh and Zambian data sets respectively have  $H12 > H12_o$ , where  $H12_o = 0.0182$  and  $0.0152$  in the Raleigh and Zambian data sets, respectively. Note that these  $H12_o$  values are comparable to the median values observed in the data, as we observed previously (GARUD *et al.* 2015).

### ***Identification of selective sweeps in the Zambian and Raleigh data***

To identify individual selective sweeps, we performed an H12 scan along the four autosomal arms of *D. melanogaster* in the less stringently filtered Zambian and Raleigh data sets (Figure 2). We used the less stringently filtered data sets to avoid missing sweeps due to the extensive masking in the stringently filtered data. Before performing the scan, we compared the distributions of H12 values measured in the stringently and less-stringently masked data sets and found that the distributions are similar (Figure S3, Figures 1c and 1d) with an elevation of the tails and medians of H12 values in the data relative to neutral demographic simulations. We

applied H12 to Raleigh data in windows of 401 SNPs and to Zambian data in windows of 801 SNPs randomly down sampled once to 401 SNPs (Figure 2).

To identify individual selective sweeps in the Zambian and Raleigh data, we called H12 peaks, where a peak is defined as a consecutive tract of H12 windows lying above the median H12 value in the data (GARUD *et al.* 2015) (Text S1). To focus on the most extreme selected events, we highlighted the top 25 H12 peaks in each scan in Figure 2. As previously observed in Figures 1c, 1d and S3, the H12 values at the peaks in the Zambian data are lower overall than in the Raleigh data. Reflecting this lower level of H12 in Zambian data, in Figure 2 the Zambian scan has a shorter y-axis than the Raleigh scan to have better resolution to identify individual selection events, but we show this data again in Figure S4 with the same y-axis as in Raleigh. In Table S4, we give the start and end coordinates of the top 25 peaks found in the Zambian and Raleigh data sets.

H12 peaks are distributed across the autosomal arms in both the Zambian and Raleigh populations, suggesting that no single region of the genome is driving the elevation of H12 values observed in the data. The exception to this is an enrichment of peaks on chromosome 3R where there are inversions present at high frequencies in Zambian data (Figure 2) (see Table S5 for the enrichment calculations).

The overlap in peaks in the Zambian and Raleigh scans in Figure 2 is small but significant. Exactly three of the top 25 peaks in the Raleigh scan overlap, which is greater than expected by chance (one sided binomial probability  $< 0.022$ , see Text S1 for calculations). One of the overlapping peaks coincides with the positive control *Cyp6g1* on chromosome 2R, and two of the overlapping peaks are on chromosome 3R.

The lack of coincidence at the rest of peaks does not necessarily mean that the selective events in Zambia and Raleigh are independent; in cases where there is no overlap, it is possible that there are sweeps in one of the populations that did not rank among the top 25 peaks. For example we recovered H12 peaks at the positive controls *Cyp6g1* and *Ace* in both the Raleigh and Zambian data sets. While peaks at *Cyp6g1* and *Ace* rank 1<sup>st</sup> and 3<sup>rd</sup> in the Raleigh scan, they rank 14<sup>th</sup> and 35<sup>th</sup> in the Zambian scan. Therefore *Ace* does not show up as an overlap among the top 25 peaks. The lower rank of the peaks at *Ace* in the Zambian scan reflects the lower frequencies at which the adaptive mutations at *Ace* are present in the Zambian and Raleigh data respectively: approximately 16% in the Zambian and approximately 40% in the Raleigh

populations. The second-ranking peak in the Raleigh data is at the positive control *CHKov1*, reiterating our previous result (GARUD *et al.* 2015). However, this peak is absent in the Zambian scan, consistent with the previous observation that the adaptive mutation at *CHKov1* is present at high frequencies in out-of-Africa populations and rare in Africa (AMINETZACH *et al.* 2005).

To determine if the peaks in Zambia are truly independent from those in Raleigh, we investigated whether any of the top three common haplotypes among the 25 peaks in the Zambian data are also common in Raleigh. We focused on the top three haplotypes because these comprise the bulk of the haplotype homozygosity, H12, at the top peaks. To identify shared common haplotypes, we examined haplotype frequency spectra in a joint data set consisting of both Zambian and Raleigh lines. We looked for matching haplotypes in the analysis windows matching the start and end coordinate positions of the analysis windows with the highest H12 values in each of the 25 Zambian peaks. We found only one peak with a common haplotype shared among 9 Zambian individuals and 8 Raleigh individuals. This peak is at the location of the *Cyp6g1* selective sweep. We did the reverse investigated whether any of the top three common haplotypes among the 25 peaks in the Raleigh data are also common in Zambia. We found substantial haplotype sharing at *Ace* with 7 Zambian haplotypes matching 34 Raleigh haplotypes. The peak closest to *Ace* had 3 Zambian haplotypes matching 7 Raleigh haplotypes, potentially because this peak belongs to the same selective event as *Ace*. At *CHKov1*, 8 Zambian haplotypes matched 34 Raleigh haplotypes. We did not find any significant haplotype sharing at *Cyp6g1*. This may be because the centers of the peaks at *Cyp6g1* in the Zambian and Raleigh populations are slightly offset.

### ***Haplotype structure at the top H12 peaks in the Zambian and Raleigh scans***

To gain intuition about the haplotype structure at the top peaks in the H12 scans of the Raleigh and Zambian data sets, we visualized the haplotype frequency spectra in the analysis window with the highest H12 value at each of the top 25 peaks (Figure 3). There is substantial haplotype homozygosity in the two populations, which we previously saw cannot be generated by the simple neutral demographic scenarios tested in Figure 1.

The signals in Zambian data are weaker than in Raleigh with the sweeps reaching a smaller partial frequency. This is consistent with the muted elevation in linkage disequilibrium and haplotype homozygosity in the Zambian data set relative to the Raleigh data set. In the



Zambian data, some peaks have a single dominant haplotype, which is consistent with signatures of hard sweeps, while many peaks have multiple frequent haplotypes, consistent with signatures of soft sweeps. This is unlike Raleigh data where multiple high frequency haplotypes appear at all the top peaks, as we previously observed (GARUD *et al.* 2015).

In the next section, we will consider several scenarios that can generate the haplotype structure we observe in the data including (i) more complex demographic processes than we have already tested, such as undetected admixture and backflow, (ii) hard sweeps that have decayed early in the sweep trajectory (MESSER AND NEHER 2012; SCHRIDER *et al.* 2015), and (iii) soft sweeps.

### ***Could demographic processes generate the bulk of the haplotype structure in Zambian data?***

We showed earlier in Figure 1 that simple neutral demographic models including bottleneck models, bottlegrowth models, and constant  $N_e$  models cannot generate the elevated LD and haplotype homozygosity observed in the data. However, it is possible that more complex effects that are not captured by the tested demographic models may be contributing to the elevated H12 values in the data. The demographic processes we test in this section include structure, admixture and backflow of individuals from North America to Zambia.

We tested for any underlying population structure in the Zambian data. To do so, we tested whether the same strains show up more often than others among the top three haplotypes at each peak. We focused on the top three haplotypes because they comprise the bulk of the haplotype homozygosity measured by H12. Almost all the Zambian strains (186 out of 188) appear at least once among the top three haplotypes across all 25 peaks. However, some strains are found in the top haplotypes more often than others. There is a slight enrichment of strains appearing in the top haplotypes: 15 strains show up in the top 3 haplotypes in 9 or more peaks, but not necessarily in the same peak (P-value = 0.0017, see Text S1 for permutation analysis). This suggests that some of the long-range linkage disequilibrium observed in Figure 1 can be explained by some low level structure in the data, but that the majority of the haplotype homozygosity at the peaks is not driven by the same strains.

Undetected backflow of flies from North America to Zambia could result in common haplotypes in North America migrating to Zambia, which could generate H12 peaks in the Zambian data. Thus we tested whether the bulk of the homozygosity at the top 25 peaks in the

Zambian data set can be explained by backflow from North America to Zambia. We compared genome-wide admixture levels in strains comprising the 3 most common haplotypes versus strains comprising all other haplotypes in the analysis window with the highest H12 value at each of the 25 peaks. We used admixture tracts inferred by POOL *et al.* (2012) for each Zambian strain for our analysis. We found that of the 25 peaks tested, only 1 peak, located on chromosome 3R, had significant differences in levels of admixture (P-value = 0.01743997, Wilcoxon two sided test), consistent with the expected number of false positives in 25 tests with a 0.05 significance level.

Backflow of flies from North America to Zambia could result in the sampling of genomic regions of low diversity ( $\pi$ ) since the haplotypes in these regions might be at greater frequencies in North America. As a consequence, these regions could generate H12 peaks in Zambia. We tested this scenario by measuring diversity ( $\pi$ ) in the less stringently filtered Raleigh data set in regions corresponding to the locations of top 25 peaks in the Zambian data versus all other regions in the Raleigh data. We found that the mean  $\pi$  in the two types of regions are similar (P-value = 0.5393864, see Text S1), which indicates that backflow is not responsible for generating the bulk of the haplotype structure at the top 25 peaks.

Finally, we also tested the reverse scenario of whether H12 peaks identified in the Raleigh data are enriched in ‘ancestry disequilibrium hubs’ identified by POOL 2015. An ancestry disequilibrium hub is characterized as a cluster of several African-African or European-European pairs of SNP with higher than expected levels of LD, suggesting that African-European SNP pairs are less compatible. We found that 2 of the top 25 peaks identified in the Raleigh scan overlap ancestry disequilibrium hubs (P-value: 0.2820319, binomial, see methods), which is not different from what is expected by chance.

Given the lack of evidence for admixture or backflow driving the elevation in haplotype homozygosity at the top peaks in the Zambian data, it is likely that the majority of H12 peaks observed in the Raleigh and Zambian data sets are driven by independent forces in the two populations, with the exception of the positive controls, as we previously saw and will further investigate below.

### ***Haplotype sharing in the Zambian and Raleigh populations at *Ace*, *Cyp6g1*, and *CHKov1****

In Figure 4 we further examine the haplotype frequency spectra at the *Ace*, *Cyp6g1* and *CHKov1* loci in a joint data set consisting of both *Zambian* and *Raleigh* lines to determine the extent of haplotype sharing at these loci. The population each haplotype is from is indicated. In this figure, the analysis window is centered around the known center of the selective sweep instead of the start and end positions of the analysis windows with the highest H12 values the selective sweep scans.

At *Ace*, a common haplotype bearing three adaptive mutations previously discovered to confer resistance to pesticides in both *Zambia* and *Raleigh* is shared between 14 strains from *Zambia* and 39 strains from *Raleigh*. Another common haplotype in the *Raleigh* data comprised of 17 strains matches 1 *Zambian* strain and also bears the three adaptive mutations. The sharing of these haplotypes support the finding by KARASOV *et al.* 2010 where it was observed that North American and Australian *D. melanogaster* strains share a haplotype bearing the three adaptive mutations, suggesting an out-of-Africa origin for this haplotype. Our finding provides a more complete adaptation history of *Ace*. All remaining high frequency haplotypes are comprised exclusively of either *Raleigh* or *Zambian* strains, many of which bear one or three adaptive mutations, suggesting that this soft sweep has arisen multiple times on multiple continents.

At the *CHKoVI* locus, we observe two common *Raleigh* haplotypes matching rare *Zambian* haplotypes: 17 *Raleigh* individuals and 2 *Zambian* individuals share one haplotype and 12 *Raleigh* individuals and 1 *Zambian* individual share another haplotype. Additionally, a third haplotype is shared between 1 *Raleigh* individual and 2 *Zambian* individuals. The sharing of common *Raleigh* haplotypes with rare *Zambian* haplotypes is consistent with the result in AMINETZACH *et al.* 2005 that the adaptive allele at *CHKoVI* is rare in African populations and common in derived populations.

At *Cyp6g1*, we observe a common haplotype shared in 10 *Raleigh* individuals and 7 *Zambian* individuals. All other haplotypes are comprised exclusively of *Raleigh* or *Zambian* strains. This result suggests that this sweep has either arisen multiple times on different continents, and/or may share a common origin, similar to the sweep at *Ace*.

### ***Bayesian analysis of the softness of the peaks***

The simple demographic models and the admixture and backflow scenarios tested previously in this paper do not provide an obvious explanation for the elevated haplotype homozygosity observed in the Zambian and Raleigh data. In this section, we test whether selective sweeps can generate the patterns observed in the data.

To assess whether the top peaks in the two scans can be more easily generated under simulated hard and soft sweeps, we measured H2/H1 (GARUD *et al.* 2015) in the windows with the highest H12 values at each of the top peaks. H2/H1 is a ratio of haplotype homozygosity calculated without (H2) and with (H1) the most frequent haplotype in the sample. This statistic, used in conjunction with H12, can distinguish hard and soft sweeps as long as the sweeps are not too weak, old, or soft (GARUD *et al.* 2015). As we previously showed (GARUD *et al.* 2015), a sweep that is too weak or old cannot be distinguished from neutrality with H12 because recombination events have decayed the sweep. Similarly, a sweep that is too soft to be detected by H12 has such a high number of unique haplotypes bearing an adaptive mutation that the sweep also cannot be distinguished easily from neutrality (GARUD *et al.* 2015).

H2/H1's ability to differentiate hard and soft sweeps is dependent on the value of H12 measured in the same analysis window. We previously demonstrated (GARUD and ROSENBERG 2015) that the range of possible H2/H1 values varies inversely with H12, and therefore the two statistics must be used in conjunction with each other. H2/H1 has high power to differentiate hard and soft sweeps when H12 is large and there is strong evidence for a sweep in the first place (GARUD *et al.* 2015). Therefore, we measured H2/H1 at the top 25 peaks in the window with the highest H12 value where there is greatest evidence for a sweep.

We performed a Bayes factor analysis to assess the likelihood of a hard versus soft sweep generating a particular pair of H12 and H2/H1 values for each of the top peaks. Bayes factors are calculated as follows:  $BF = P(H12_{obs}, H2_{obs}/H1_{obs} | \text{Soft Sweep})/P(H12_{obs}, H2_{obs}/H1_{obs} | \text{Hard Sweep})$ . We approximated BFs using an approximate Bayesian computation (ABC) approach similar to our previous analysis (GARUD *et al.* 2015) where the parameters selection coefficient ( $s$ ), partial frequency of the adaptive allele after selection has ceased ( $PF$ ), and age ( $T_E$ ) are drawn from uniform prior distributions:  $s \sim U[0,1]$ ,  $PF \sim U[0,1]$ , and  $T_E \sim U[0,1000 \text{ generations}]$  (Methods). Similar to our previous analysis (GARUD *et al.* 2015), we simulated hard sweeps with an adaptive mutation rate  $\theta_A = 0.01$  and soft sweeps with  $\theta_A = 10$ . To assess the softness of the top 25 peaks in the Zambian data, we performed simulations with a constant  $N_e = 4.8 * 10^6$

demographic model (matching the inferred  $N_e$  for the Zambian population, Methods) and measured H12 and H2/H1 at the sweep center in windows of 801 SNPs down sampled to 401 SNPs. To assess the softness of the top 25 peaks in the Raleigh data, we performed simulations with a constant  $N_e=2.6 \times 10^6$  demographic model (matching the inferred  $N_e$  for the Raleigh population, Methods) and measured H12 and H2/H1 at the sweep center in windows of 401 SNPs. We performed simulations with recombination rates of 0.5 cM/Mb and 5 cM/Mb, reflecting the range of recombination rates observed in the at the top 25 peaks in Zambian and Raleigh data (0.54 to 6.7 cM/Mb).

In Figure 5 we present panels of Bayes factors measuring the relative support of hard versus soft sweeps for a grid of different values of H12 and H2/H1 for the Zambian and Raleigh data. We tested two recombination rates for our analyses: in Figures 5A and B we use  $\rho=0.5$  cM/Mb and 5 cM/Mb, respectively. We measured H12 in a window centered at the selective site. As we already showed, the H12 values measured in Zambian data are lower relative to Raleigh. Given that signals are muted in Zambian data, for some peaks it is difficult to distinguish whether the sweeps are hard or soft. Nonetheless, the analysis shows that when  $\rho=0.5$  cM/Mb, 13 of the top 25 peaks in the Zambian data have H12 and H2/H1 values that are supported by soft sweeps, 10 peaks have H12 and H2/H1 values that are supported by hard sweeps, and 2 peaks which are borderline supported by hard or soft sweeps. This is in contrast to Raleigh data which all have high H12 and H2/H1 values that are well supported by soft sweeps, as we have seen previously (GARUD *et al.* 2015). When  $\rho=5$  cM/Mb, we find that 18 of the top 25 peaks in the Zambian data have H12 and H2/H1 values that are supported by soft sweeps, 5 peaks have H12 and H2/H1 values that are supported by hard sweeps, and 2 peaks which are borderline supported by hard or soft sweeps. Again, all peaks in the Raleigh data are well supported by soft sweeps.

In Figure 5C, we tested whether measuring H12 and H2/H1 at the sweep center in simulations is anti-conservative for detecting soft sweeps. SCHRIDER *et al.* 2015 showed that regions with the highest LD measured with the statistic  $iHS$  (VOIGHT *et al.* 2006) are offset from the sweep center. Furthermore, SCHRIDER *et al.* 2015 have suggested that this may be problematic for differentiating hard and soft sweeps since ‘soft shoulders’ of hard sweeps may not be distinguishable from soft sweeps. Soft shoulders are soft-looking sweeps to the left and right of hard sweeps that have decayed from recombination and mutation events. In Figure 5C

we repeated our analysis in Figure 5A, except instead of centering our analysis window at the sweep center, we first identified the analysis window with the highest H12 value in a 100Kb simulated chromosome and then measured H2/H1 in this window. We found that our results do not change at all from Figure 5A for Zambian data. However, for Raleigh data, it is difficult to distinguish whether 4 peaks are best supported by hard or soft sweeps as they are borderline supported by both hard and soft sweeps since both can generate the observed H12 and H2/H1 values in the data.

Overall the substantial number of peaks with H12 and H2/H1 values consistent with soft sweeps in both populations suggest that soft sweeps are in general common in *D. melanogaster*.

## DISCUSSION

We previously showed that patterns of elevated LD and haplotype homozygosity consistent with soft sweeps are common in *D. melanogaster* strains from Raleigh (GARUD *et al.* 2015). Here we compared these patterns with that of a second population from Zambia to test the extent to which this signal in the Raleigh population may be driven by peculiarities of the data such as the extensive admixture of the Raleigh population or the inbreeding performed to generate the sequenced lines. Unlike the Raleigh population, the Zambian population has undergone relatively little admixture, and instead of inbreeding lines, the data set was generated by sequencing haploid embryos.

We found that both LD and haplotype homozygosity were elevated at all distances in the Zambian and Raleigh data, relative to expectations in several neutral demographic models that match  $S$  and  $\pi$  estimates in the data. This elevation remained despite controlling for inversions, admixture, regions of low recombination, relatedness between individuals, missing data, sample size, and allele frequency. We conducted a scan for selective sweeps with the haplotype homozygosity statistic H12 and found signatures of multiple haplotypes at high frequency at the top 25 peaks in both the Zambian and Raleigh data sets. We did not find evidence for admixture or backflow generating the haplotype structure observed in Zambian data. Instead, we found that the signatures observed in this population are independent of the Raleigh population, with the exception of specific cases at the positive controls, *Ace*, *Cyp6g1*, and *CHKov1*.

We performed a Bayes factor analysis to test whether the haplotype structure observed at the top 25 peaks can be generated by hard or soft sweeps. We used H12 and H2/H1 jointly to

distinguish hard and soft sweeps since  $H2/H1$  has power only when  $H12$  is large and distinguishable from neutrality (GARUD *et al.* 2015). We found that the  $H12$  and  $H2/H1$  values observed that the top 25 peaks in the Zambian and data set are supported by hard and soft sweeps and that all of the top 25 peaks in the Raleigh data set are supported by soft sweeps, consistent with our previous observation that soft sweeps are common in Raleigh (GARUD *et al.* 2015).

Why do we see so few hard sweeps, especially in Raleigh? One possibility is that hard sweeps are driven by weaker selection than we can detect with large  $\sim 10\text{Kb}$   $H12$  analysis windows. To hone our intuitions about the range of  $H12$  and  $H2/H1$  values that we can expect for sweeps of varying selective strength, in Figures 6 and S6 we measured  $H12$  and  $H2/H1$  in simulated partial hard and soft sweeps of varying selection coefficients ranging from  $10^{-6}$  to 1 with  $\rho = 0.5$  and 5 cM/Mb, reflecting the range of recombination rates observed at the top peaks in Zambian and Raleigh data (0.54 to 6.7 cM/Mb). We find that hard and soft sweeps must experience strong selection in order to be differentiable from neutrality. We also find that  $H12$  has greater power to detect hard sweeps than soft sweeps, especially for strong selection coefficients. Strong selection generates high  $H12$  values and low  $H2/H1$  values for hard sweeps. For weaker selection, it is possible to generate high  $H2/H1$  for hard sweeps, but then  $H12$  values are muted. In general, hard sweeps do not naturally generate both high  $H12$  and  $H2/H1$ , although weak selection can generate peaks with significant values of  $H12$  and  $H2/H1$ . Indeed, in the Zambian data, we do observe several peaks with low  $H12$  values and high  $H2/H1$  values; these peaks could potentially be hard sweeps that have experienced weak selection and these are the peaks for which our Bayesian analysis fails to distinguish between soft and hard sweeps. However, many peaks in the Raleigh and Zambian data have very high  $H12$  and  $H2/H1$  values and these peaks are much more easily explained by soft sweeps driven by very strong selection.

Could hard sweeps that have had a chance to decay for varying amounts of time generate soft-looking sweeps? In Figures 7 and S7 we measured the range of  $H12$  and  $H2/H1$  values we can expect for strong hard and soft sweeps that have decayed for varying amounts of time. We focused on strong sweeps in this example because in Figures 6 and S6 we showed that weak sweeps could not be detected easily even when there has not been any decay. We find in Figures 7 and S7 that hard sweeps that have the opportunity to decay for a specific amount of time can generate high  $H2/H1$  values, but the  $H12$  values in these cases are extremely muted and are only

borderline higher than values expected under neutrality. We also find that decay eliminates the ability to detect soft sweeps. Thus, the soft sweeps we find in the data are likely to be very young.

Could the sides of hard sweeps, which have decayed due to recombination events, generate soft-looking sweeps, or “soft shoulders” (SCHRIDER *et al.* 2015)? In Figures 8 and S10, we measured H12 and H2/H1 values in windows at varying distances away from the center of hard sweeps where the selected site is located. Again in this example we focused on strong sweeps because we showed in Figures 6 and S6 that sweeps could not be detected easily when experiencing weak selection. We found that while H2/H1 values do increase to some extent further away from the sweep center, H12 values are the highest at the selected site where H2/H1 values are also low. This highlights the importance of first detecting a sweep with an H12 value that is significantly differentiable from neutrality before testing whether it is hard or soft. Additionally, within a peak, it is crucial to first identify the analysis window with the highest H12 value in a peak before applying H2/H1; otherwise, there is a risk of misclassifying hard sweeps as soft. Note that in our analysis, we define a peak as a consecutive tract of analysis windows lying above the median H12 value in the data, which as we showed previously, is similar to the one-per-genome FDR value for H12 under neutrality (GARUD *et al.* 2015).

The extent of the elevation in LD and haplotype homozygosity and the number of soft sweeps in both populations is surprising and simultaneously confirms that these results are generic to multiple populations of *D. melanogaster*. Yet, there are some differences in the patterns observed in the Zambian and Raleigh data set. For example, the elevation in LD and haplotype homozygosity is muted in the Zambian population relative to the Raleigh population and all the selective sweeps in Zambia have lower partial frequencies and H12 values than in Raleigh. In addition, while many of the top 25 peaks’ H12 and H2/H1 values in the Zambian data are strongly supported by soft sweeps, several top peaks are consistent with hard sweeps driven by weaker selection. Note that it is difficult to differentiate whether these sweeps are truly hard or soft because H2/H1 has poor ability to differentiate the softness of a sweep with a low H12 value (GARUD *et al.* 2015).

Much stronger signals of soft sweeps in the Raleigh data set could in principle be driven by the inbreeding and admixture experienced by the Raleigh population, but it is clear that these two forces alone cannot explain all the elevation in haplotype homozygosity and LD given that these signatures are generic to both populations. Instead, another possibility why the signals in



the Raleigh population are more pronounced could be that Raleigh has experienced stronger and more recent selection pressures, especially due to anthropogenic pressures such as those arising from the presence of pesticides or due to novel selective pressures associated with migration to novel temperate habitats. The patterns at the positive controls *Ace*, *Cyp6g1*, and *CHKov1*, which are cases of recent and strong positive selection arising from the presence of pesticides and viruses are consistent with this interpretation. The frequencies of the adaptive mutations at the positive control *Ace* shows that the adaptive mutations have reached frequencies of ~40% in Raleigh and only ~16% in Zambia. The peaks at *Cyp6g1* and *Ace* rank 14<sup>th</sup> and 35<sup>th</sup> in the H12 scan in Zambia, in contrast to 1<sup>st</sup> and 3<sup>rd</sup> in the Raleigh scan likely because the adaptive mutations have reached lower frequencies in Zambia. We do not even observe a peak at *CHKov1* in Zambia, consistent with the finding in AMINETZACH *et al.* 2005 that the adaptive allele is rare in Africa and common only in derived populations at this locus.

At all three loci, *Ace*, *CHKov1*, and *Cyp6g1* multiple haplotypes at high frequencies are unique to both the Zambian and Raleigh populations, showing that independent soft sweeps are common in different locations around the world. Additionally, at all three loci, we observed that a common haplotype is shared by both Zambia and Raleigh strains, suggesting that migration also plays a role. At *Ace*, the shared common haplotype bears three adaptive mutations in contrast to haplotypes bearing the single adaptive mutation present at high frequencies comprised uniquely of Zambia or Raleigh strains. This suggests that this haplotype arose once in Africa and spread worldwide, confirming the results of KARASOV *et al.* 2010. At *Cyp6g1* we find for the first time that a common haplotype is shared in both Zambia and Raleigh. At *CHKov1* we find that common haplotypes in Raleigh are shared with rare haplotypes in Zambia, consistent with the result in AMINETZACH *et al.* 2005 showing that the adaptive allele originated in Africa and rose to high frequencies in derived populations.

Intriguingly, we notice that virtually all sweeps in Raleigh and Zambia reach an intermediate frequency and none go to completion. It is possible that most adaptation is local or environment-specific or that most adaptive alleles are only adaptive when rare and driven by frequency-dependent selection and then maintained at intermediate frequencies by balancing selection. The prevalence of incomplete sweeps may have gone unnoticed until now because previous studies often relied on shallower sample sizes in which sweeps may have looked like fixations due to sampling or could not be distinguished from neutrality. Current McDonald-

Kreitman estimates of the rate of adaptive mutations entering the population (ANDOLFATTO 2005; BIERNE AND EYRE-WALKER 2004; EYRE-WALKER AND KEIGHTLEY 2009; FAY *et al.* 2002; KOUSATHANAS AND KEIGHTLEY 2013; SCHNEIDER *et al.* 2011; SELLA *et al.* 2009; SHAPIRO *et al.* 2007; SMITH AND EYRE-WALKER 2002) may be underestimates if partial sweeps are common and often fail to go to fixation in the long term because McDonald-Kreitman estimates measure rate of adaptation by focusing on the rate of fixation over long time scales. This will need to be tested further with forward simulations of realistic demographic and selection scenarios.

Although we tested several neutral demographic scenarios fit to the Zambian and Raleigh data, we could not match the overall elevation in LD and haplotype homozygosity observed in the data. It is possible that there is a demographic scenario yet to be identified that can generate the observed levels of LD and haplotype homozygosity. One such extreme scenario that could generate the patterns observed is an admixture model where two parental populations undergo sharp bottlenecks resulting in virtually isogenic populations followed by an expansion before mixing. Alternatively, there could be demes within each population generating structure in the data that we are unable to model. Going forward, simulations of more realistic demographic scenarios potentially also incorporating selection are needed to fully understand the underlying forces in these populations.

Although background selection and gene conversion act pervasively throughout the *D. melanogaster* genome (COMERON 2014; COMERON *et al.* 2012), it is unlikely that these forces significantly influence the patterns of LD and haplotype homozygosity: while background selection can depress levels of polymorphisms, on its own it cannot elevate haplotypes to high frequencies (ENARD *et al.* 2014; FAGNY *et al.* 2014). Gene conversion has been shown to act on short scales of 50 to 2000 bps at most (HILLIKER *et al.* 1994; JEFFREYS AND MAY 2004) and should thus only impact LD on short scales and not long-range LD. In fact, gene conversion should decrease LD on short scales, thereby making haplotype homozygosity tests that do not take into account gene conversion conservative (ANDOLFATTO AND NORDBORG 1998).

The nature of selective or demographic forces that generate such high levels of LD in *D. melanogaster* is puzzling and remains to be elucidated. Likewise, we do not yet understand the nature of selection that generates such abundant partial strong soft sweeps and weaker hard sweeps. High levels of long range LD has been observed in many other organisms in addition to *D. melanogaster* including cattle, wheat, maize, sheep, and humans (FARNIR *et al.* 2000;

MACCAFERRI *et al.* 2005; MILLER *et al.* 2011; PALAISA *et al.* 2004; PRITCHARD AND PRZEWORSKI 2001; RYBICKI *et al.* 2002). Thus the evolutionary forces explored in this study and future analysis of LD and haplotype homozygosity in *D. melanogaster* should prove illuminating for other species as well.

## ACKNOWLEDGEMENTS

We gratefully thank Philipp Messer, David Enard, Noah Rosenberg, Shaila Musharoff, Alan Bergland, Emily Ebel, David Lawrie, Arbel Harpak, and Gary Howard for their invaluable help. This work was supported by the Stanford Center for Evolution and Human Genomics fellowship to NRG and NIH grants R01 GM100366, R01 GM097415, R01 GM089926 to DAP. Part of this paper was written in Katie Pollard's lab at UCSF.

## REFERENCES

- ABECASIS, G. R., D. ALTSHULER, A. AUTON, L. D. BROOKS, R. M. DURBIN *et al.*, 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.
- AMINETZACH, Y. T., J. M. MACPHERSON AND D. A. PETROV, 2005 Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science* **309**: 764-767.
- ANDOLFATTO, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**: 1149-1152.
- ANDOLFATTO, P., AND M. NORDBORG, 1998 The effect of gene conversion on intralocus associations. *Genetics* **148**: 1397-1399.
- ANDOLFATTO, P., AND M. PRZEWORSKI, 2000 A Genome-Wide Departure From the Standard Neutral Model in Natural Populations of *Drosophila*. *Genetics* **156**: 257-268.
- ANDOLFATTO, P., AND J. D. WALL, 2003 Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics* **165**: 1289-1305.
- BERGLAND, A. O., R. TOBLER, J. GONZALEZ, P. S. SCHMIDT AND D. A. PETROV, 2015 Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. bioRxiv.
- BIERNE, N., AND A. EYRE-WALKER, 2004 The genomic rate of adaptive amino acid substitution in *Drosophila*. *Molecular Biology and Evolution* **21**: 1350-1360.
- CAO, J., K. SCHNEEBERGER, S. OSSOWSKI, T. GUNTHER, S. BENDER *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature genetics* **43**: 956-963.
- COMERON, J. M., 2014 Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genetics* **10**: e1004434.
- COMERON, J. M., R. RATNAPPAN AND S. BAILIN, 2012 The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics* **8**: e1002905.
- CORBETT-DETIG, R. B., AND D. L. HARTL, 2012 Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genetics* **8**: e1003056.

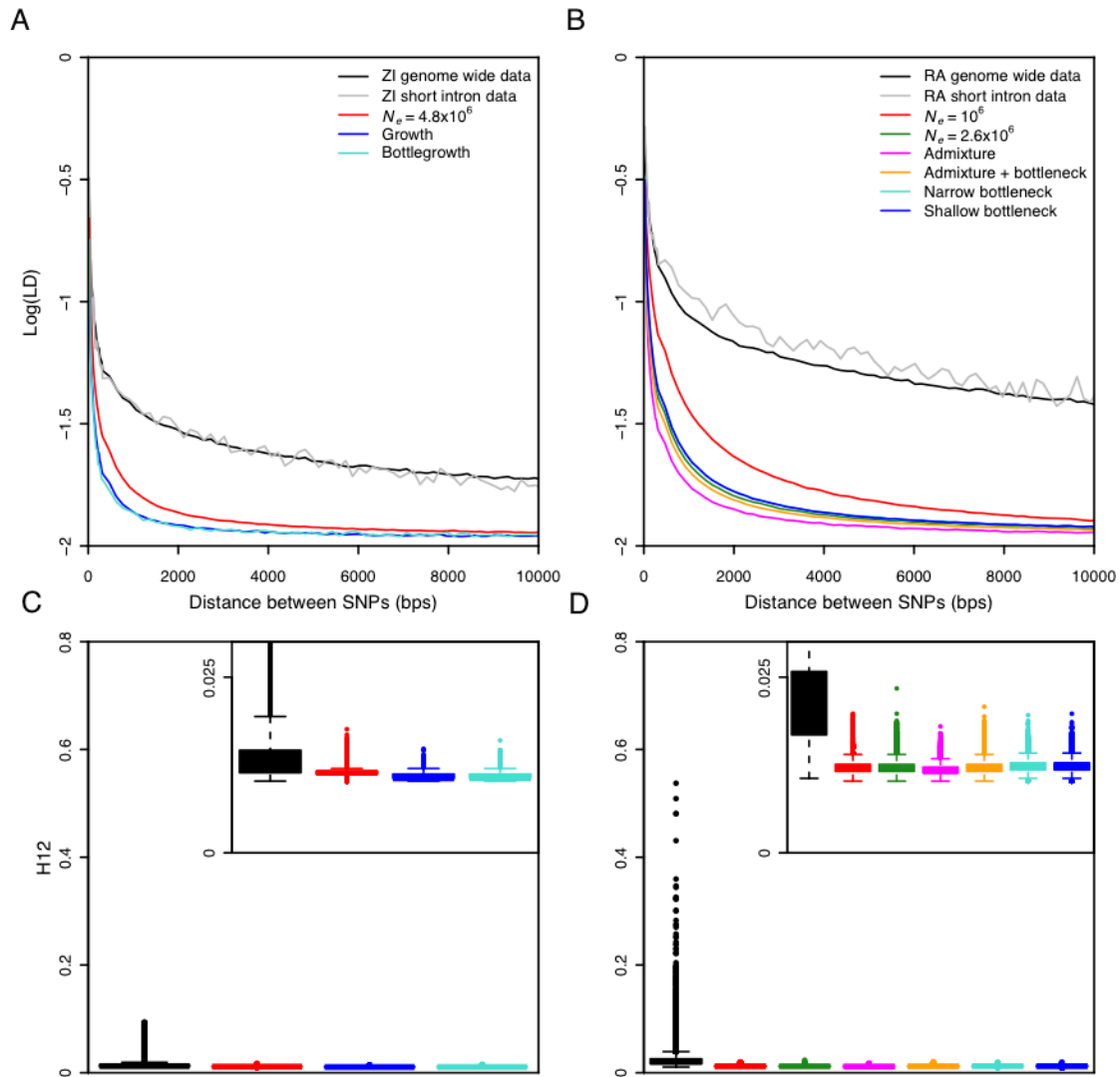
- DUCHEN, P., D. ZIVKOVIC, S. HUTTER, W. STEPHAN AND S. LAURENT, 2013 Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. *Genetics* **193**: 291-301.
- ENARD, D., P. W. MESSER AND D. A. PETROV, 2014 Genome-wide signals of positive selection in human evolution. *Genome Research* **24**: 885-895.
- EWING, G., AND J. HERMISSON, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**: 2064-2065.
- EYRE-WALKER, A., AND P. D. KEIGHTLEY, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution* **26**: 2097-2108.
- FAGNY, M., E. PATIN, D. ENARD, L. B. BARREIRO, L. QUINTANA-MURCI *et al.*, 2014 Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Molecular Biology and Evolution* **31**: 1850-1868.
- FARNIR, F., W. COPPIETERS, J. J. ARRANZ, P. BERZI, N. CAMBISANO *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. *Genome research* **10**: 220-227.
- FAY, J. C., G. J. WYCKOFF AND C. I. WU, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**: 1024-1026.
- GARUD, N. R., P. W. MESSER, E. O. BUZBAS AND D. A. PETROV, 2015 Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics* **11**: e1005004.
- GARUD, N. R., AND N. A. ROSENBERG, 2015 Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps. *Theoretical Population Biology*.
- GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON AND C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**: e1000695.
- HILLIKER, A. J., G. HARAUZ, A. G. REAUME, M. GRAY, S. H. CLARK *et al.*, 1994 Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. *Genetics* **137**: 1019-1026.
- HOULE, D., AND E. J. MARQUEZ, 2015 Linkage Disequilibrium and Inversion-Typing of the *Drosophila melanogaster* Genome Reference Panel. *G3*: 1695-1701.
- HUANG, W., A. MASSOURAS, Y. INOUE, J. PEIFFER, M. RAMIA *et al.*, 2014 Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome research* **24**: 1193-1208.
- JAKOBSSON, M., S. W. SCHOLZ, P. SCHEET, J. R. GIBBS, J. M. VANLIERE *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998-1003.
- JEFFREYS, A. J., AND C. A. MAY, 2004 Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature genetics* **36**: 151-156.
- KARASOV, T., P. W. MESSER AND D. A. PETROV, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genetics* **6**: e1000924.
- KEIGHTLEY, P. D., U. TRIVEDI, M. THOMSON, F. OLIVER, S. KUMAR *et al.*, 2009 Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research* **19**: 1195-1201.

- KOUSATHANAS, A., AND P. D. KEIGHTLEY, 2013 A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* **193**: 1197-1208.
- LACK, J. B., C. M. CARDENO, M. W. CREPEAU, W. TAYLOR, R. B. CORBETT-DETIG *et al.*, 2015 The *Drosophila* Genome Nexus: A Population Genomic Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range Population. *Genetics* **199**: 1229-1241.
- LANGLEY, C. H., M. CREPEAU, C. CARDENO, R. CORBETT-DETIG AND K. STEVENS, 2011 Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics* **188**: 239-246.
- LAWRIE, D. S., P. W. MESSER, R. HERSHBERG AND D. A. PETROV, 2013 Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics* **9**: e1003527.
- LI, H., AND W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genetics* **2**: e166.
- MACCAFERRI, M., M. C. SANGUINETI, E. NOLI AND R. TUBEROSA, 2005 Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Molecular Breeding* **15**: 1380-3743.
- MACKAY, T. F., S. RICHARDS, E. A. STONE, A. BARBADILLA, J. F. AYROLES *et al.*, 2012 The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173-178.
- MAGWIRE, M. M., F. BAYER, C. L. WEBSTER, C. CAO AND F. M. JIGGINS, 2011 Successive increases in the resistance of *Drosophila* to viral infection through a transposon insertion followed by a Duplication. *PLoS Genetics* **7**: e1002337.
- MESSER, P. W., AND R. A. NEHER, 2012 Estimating the strength of selective sweeps from deep population diversity data. *Genetics* **191**: 593-605.
- MILLER, J. M., J. POISSANT, J. W. KIJAS AND D. W. COLTMAN, 2011 A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular ecology resources* **11**: 314-322.
- MOLINA, J., M. SIKORA, N. GARUD, J. M. FLOWERS, S. RUBINSTEIN *et al.*, 2001 Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 8351-8356.
- PALAISSA, K., M. MORGANTE, S. TINGEY AND A. RAFALSKI, 2004 Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 9885-9890.
- PARSCH, J., S. NOVOZHILOV, S. S. SAMINADIN-PETER, K. M. WONG AND P. ANDOLFATTO, 2010 On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution* **27**: 1226-1234.
- PENNINGS, P. S., AND J. HERMISSON, 2006 Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genetics* **2**: e186.
- POKALYUK, C., 2012 The effect of recurrent mutation on the linkage disequilibrium under a selective sweep. *Journal of Mathematical Biology* **64**: 291-317.
- POOL, J. E., 2015 Natural Selection Shapes the Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the *D. melanogaster* Reference Genome. *bioRxiv*.
- POOL, J. E., R. B. CORBETT-DETIG, R. P. SUGINO, K. A. STEVENS, C. M. CARDENO *et al.*, 2012 Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS genetics* **8**: e1003080.

- PRITCHARD, J. K., AND M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. *American journal of human genetics* **69**: 1-14.
- PRITCHARD, J. K., M. STEPHENS AND P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.
- PRZEWORSKI, M., J. D. WALL AND P. ANDOLFATTO, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution* **18**: 291-298.
- RYBICKI, B. A., S. K. IYENGAR, T. HARRIS, R. LIPTAK, R. C. ELSTON *et al.*, 2002 The distribution of long range admixture linkage disequilibrium in an African-American population. *Human heredity* **53**: 187-196.
- SCHMIDT, J. M., R. T. GOOD, B. APPLETON, J. SHERRARD, G. C. RAYMANT *et al.*, 2010 Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genetics* **6**: e1000998.
- SCHNEIDER, A., B. CHARLESWORTH, A. EYRE-WALKER AND P. D. KEIGHTLEY, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* **189**: 1427-1437.
- SCHRIDER, D. R., F. K. MENDES, M. W. HAHN AND A. D. KERN, 2015 Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**: 267-284.
- SELLA, G., D. A. PETROV, M. PRZEWORSKI AND P. ANDOLFATTO, 2009 Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics* **5**: e1000495.
- SHAPIRO, J. A., W. HUANG, C. ZHANG, M. J. HUBISZ, J. LU *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 2271-2276.
- SMITH, N. G., AND A. EYRE-WALKER, 2002 Adaptive protein evolution in *Drosophila*. *Nature* **415**: 1022-1024.
- THORNTON, K., AND P. ANDOLFATTO, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**: 1607-1619.
- VOIGHT, B. F., S. KUDARAVALLI, X. WEN AND J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. *PLoS Biology* **4**: e72.
- WALL, J. D., AND J. K. PRITCHARD, 2003 Haplotype blocks and linkage disequilibrium in the human genome. *Nature reviews. Genetics* **4**: 587-597.

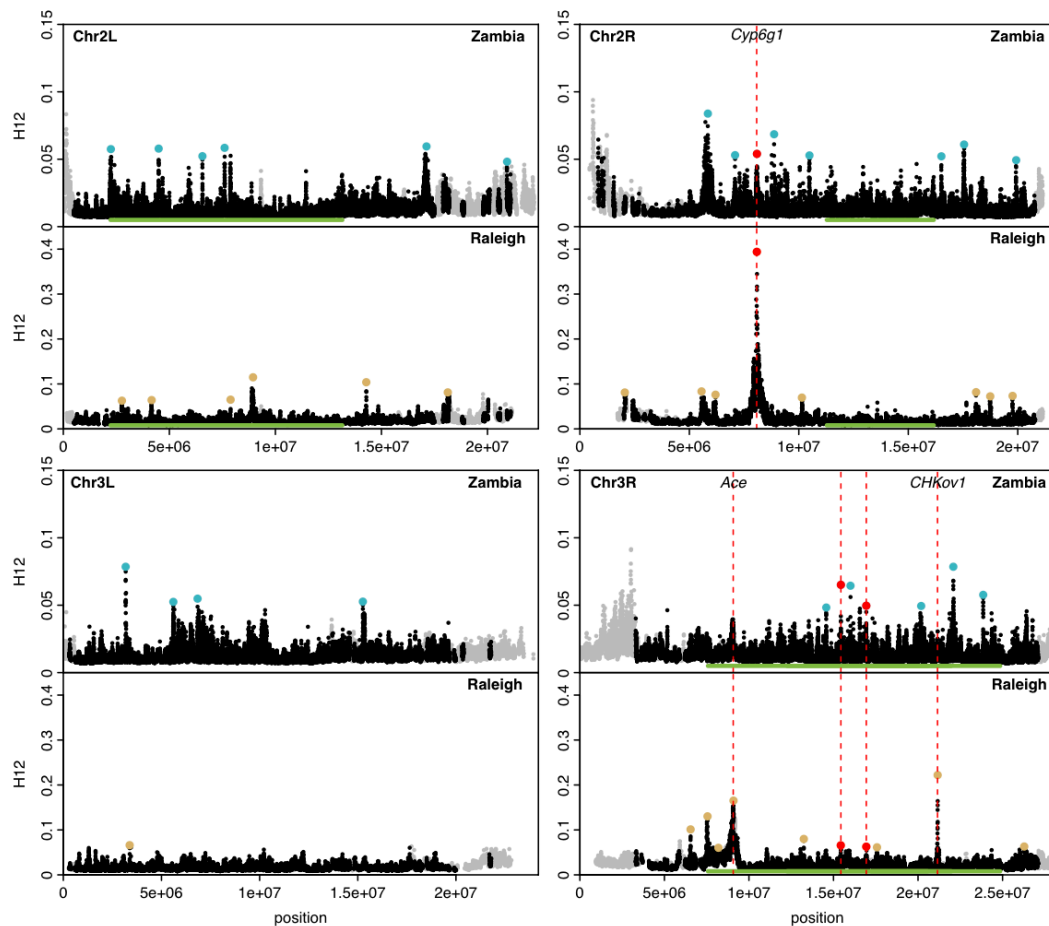
## FIGURES

**Figure 1.** Elevated long-range LD and haplotype homozygosity in Zambia and Raleigh populations. LD in Zambia (ZI) (A) and Raleigh (RA) (B) is elevated as compared to any neutral demographic model, especially for long distances. H12 values measured in 801 SNP windows down sampled to 401 SNPs in Zambia data (C) and 401 SNP windows in Raleigh data (D) were higher than expectations under any neutral demographic model tested. Pairwise LD and H12 were calculated in Zambia and Raleigh data along the regions of the autosomal arms of the *D. melanogaster* genome with  $\rho \geq 1$  cM/Mb in samples of 100. Neutral demographic simulations were generated with  $\rho = 10^{-6}$  cM/bp. In A and B, pairwise LD was averaged over  $10^7$  simulations in each neutral demographic scenario. In C and D, plotted are the result of approximately  $1.5 \times 10^5$  simulations under each neutral demographic model, representing more than 10 times the number of analysis observed windows in the data.



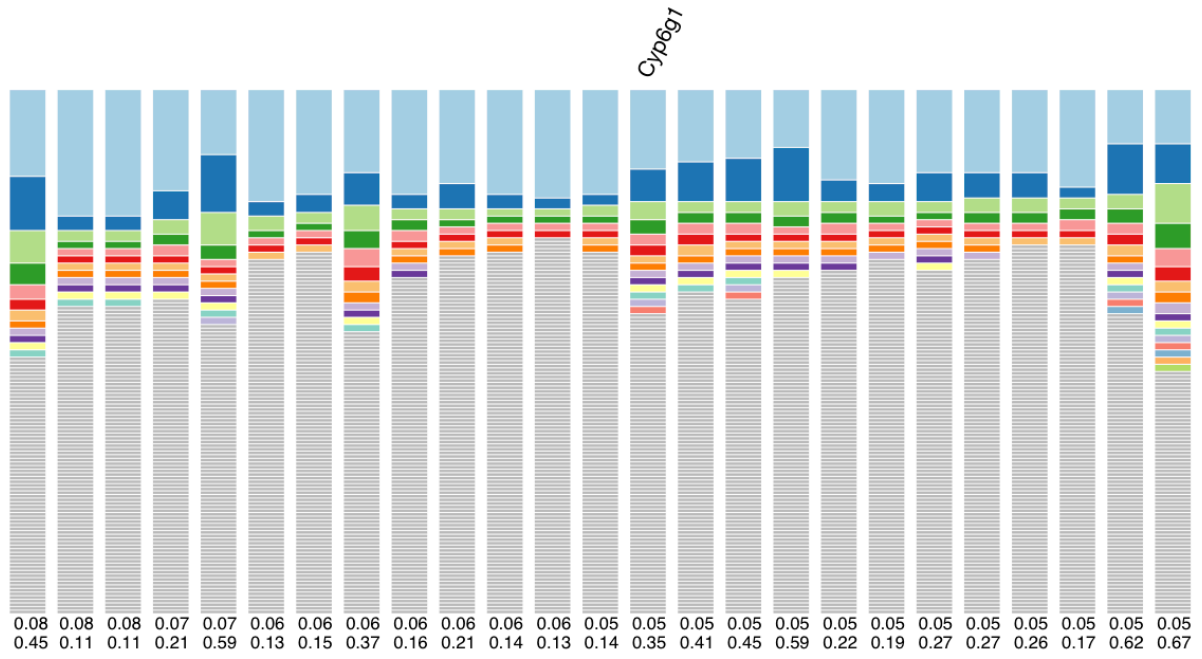


**Figure 2.** H12 scan in Zambian and Raleigh data. We performed an H12 scan in Zambian data with windows of size 801 SNPs down-sampled to 401 SNPs and in Raleigh data with windows of size 401 SNPs. Each data point represents the H12 value calculated in an analysis window centered at the particular genomic position, with window centers separated by 50 SNPs. Grey points indicate regions on the autosomal arms with recombination rates lower than 0.5 cM/Mb (COMERON *et al.* 2012) excluded from our analysis. Green bars indicate common inversions. There are multiple overlapping inversions on chromosome 3R. Red and blue points highlight the top 25 H12 peaks in the each respective scan relative to the median H12 value in the data. Red points indicate peaks that overlap both scans. The three highest peaks in the Raleigh scan correspond to known cases of soft sweeps at *Ace*, *CHKov1*, and *Cyp6g1*. We also observed peaks at *Ace* and *Cyp6g1* in the Zambian scan, but these did not rank as highly because the adaptive alleles at these loci were at lower frequencies at these sites. See Figure S5 for a higher resolution H12 scan.

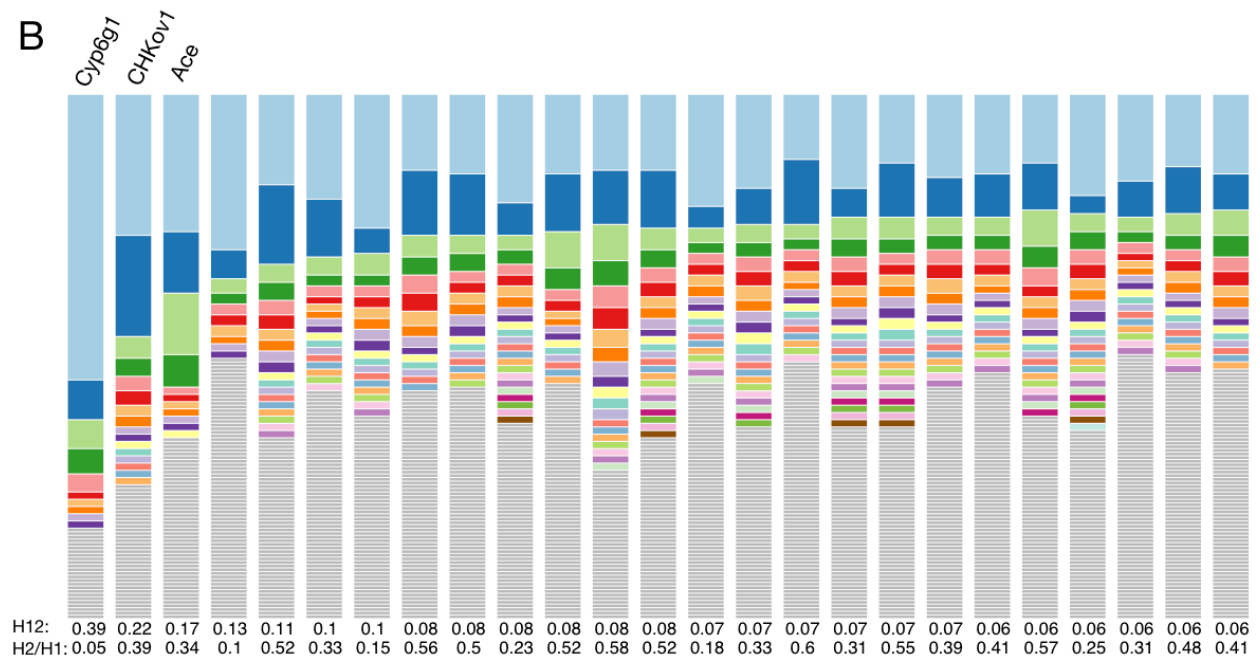


**Figure 3.** Haplotype frequency spectra for the 25 H12 peaks in Zambian and Raleigh data. Shown are haplotype frequency spectra for the top 25 peaks in the Zambian scan conducted in 801 SNP windows down-sampled to 401 SNPs (A) and the Raleigh scan conducted in 401 SNP windows (B). For each peak, the frequency spectrum corresponding to the analysis window with the highest H12 value was plotted. The height of the upmost shaded region (light blue) in each bar indicates the frequency of the most prevalent haplotype in the sample of 145 individuals, and heights of subsequent colored bars indicate the frequency of the second, third, and so on most frequent haplotypes in a sample. Grey bars indicate singletons. In Zambia, sweeps reach a smaller partial frequency than Raleigh. Many peaks in the Zambian data have multiple haplotypes present at high frequency indicative of soft sweeps, and many peaks have a single haplotype dominating the haplotype spectra, indicative of hard sweeps. In Raleigh all sweeps have multiple haplotypes at high frequency consistent with signatures of soft sweeps.

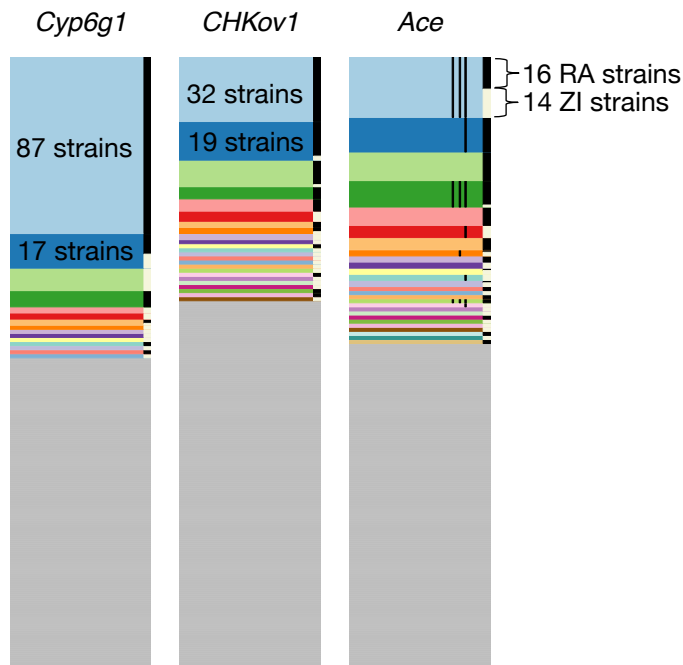
A



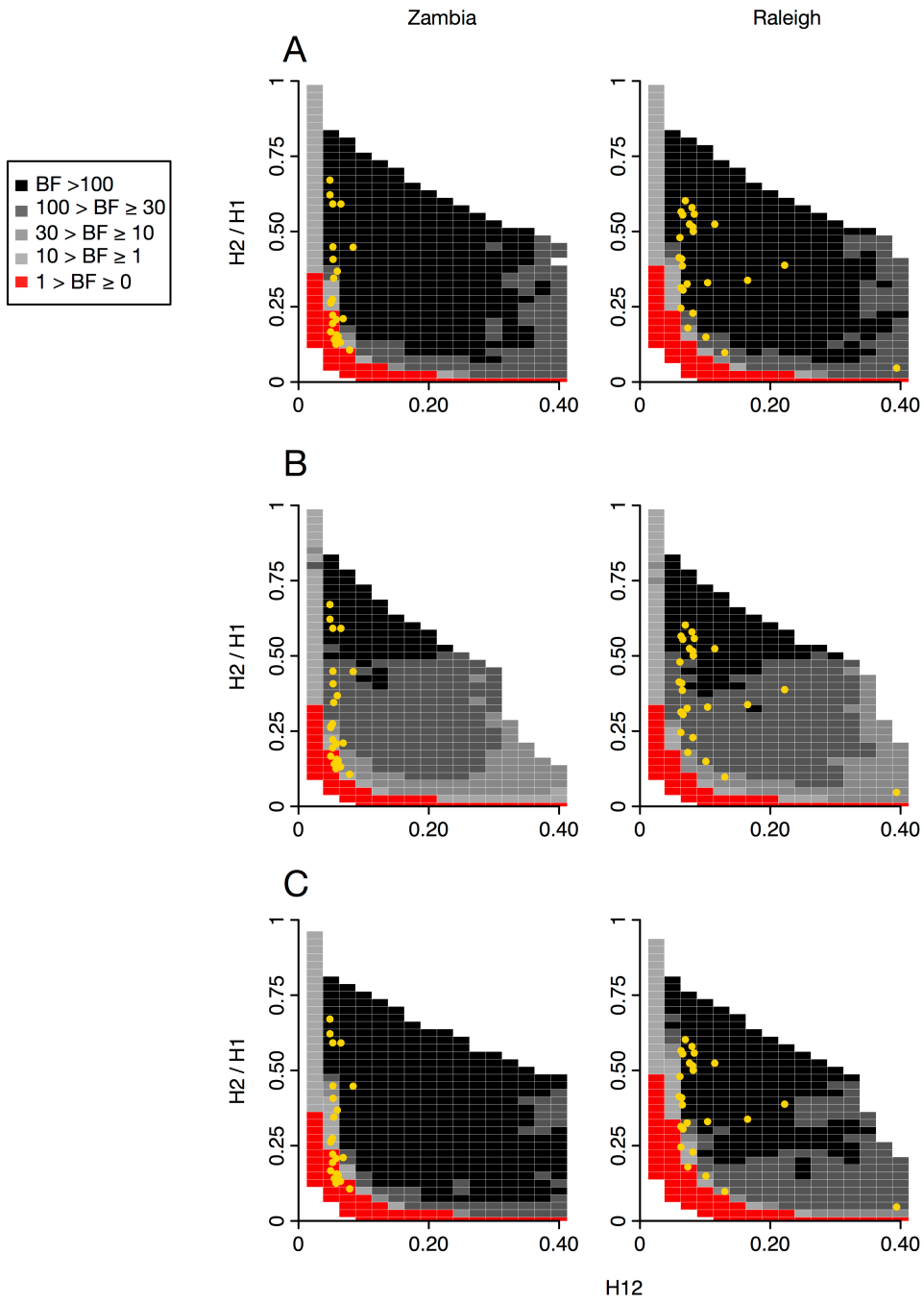
B



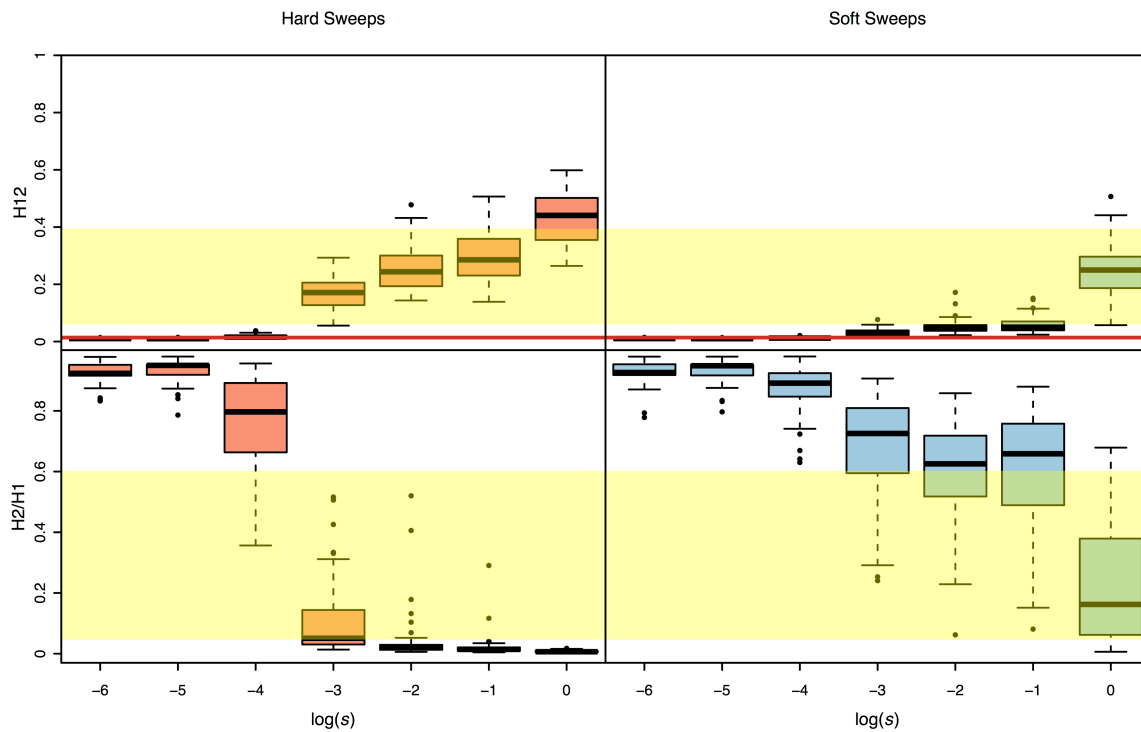
**Figure 4.** Haplotype frequency spectra at *Cyp6g1*, *CHKov1*, and *Ace*. Shown are haplotype frequency spectra at the three positive controls in a joint data set comprising 300 Raleigh (RA) strains and Zambian (ZI) strains in 801 SNP windows positioned at the sweep centers. 801 SNP windows in the joint data set correspond to a slightly smaller analysis window size (<10Kb) in terms of base pairs on average than in the Raleigh or Zambian data alone. On the right side of each of the frequency spectra are black and white bars, indicating which strains are from Raleigh (RA) and Zambia (ZI), respectively. At all three positive controls, common haplotypes are shared across the two populations. The thin black lines shown in the haplotype spectrum for *Ace* correspond to the presence of three adaptive mutations that confer resistance to pesticides at *Ace*. Three haplotypes have all three adaptive mutations. Raleigh and Zambian strains share two of these haplotypes.



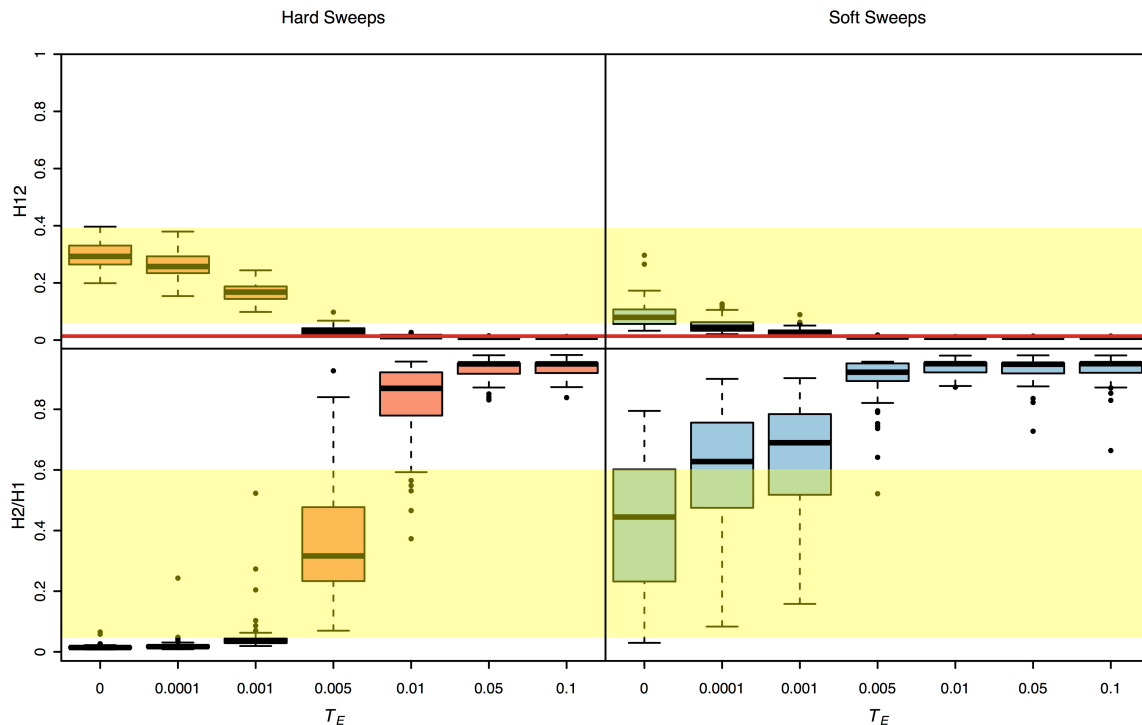
**Figure 5.** Range of H12 and H2/H1 values expected under hard and soft sweeps for Zambian and Raleigh data. Bayes factors (BFs) were calculated for a grid of H12 and H2/H1 values to demonstrate the range of H12 and H2/H1 values expected under hard versus soft sweeps for the top 25 peaks in the Zambian and Raleigh scans. BFs were calculated by taking the ratio of the number of soft sweep versus hard sweep simulations that were within a Euclidean distance of 0.1 of a given pair of H12 and H2/H1 values. Red portions of the grid represent H12 and H2/H1 values with stronger support for hard sweeps, while grey portions represent regions of space with stronger support for soft sweeps. Each panel presents the results from  $10^6$  hard sweep simulations generated with  $\theta_A = 0.01$  and  $10^6$  soft sweep simulations generated with  $\theta_A = 10$ . For any colored square, at least 1000 hard and soft sweep simulations match the corresponding pair of H12 and H2/H1 values of the grid. White portions of the grid represent cases where fewer than 1000 simulations match the pair of H12 and H2/H1 values. Simulations were performed in samples of 145 individuals with a constant  $N_e \approx 4.8 \times 10^6$  demographic model for comparison with the Zambian data and a constant  $N_e \approx 2.6 \times 10^6$  demographic model for comparison with the Raleigh data. Overlaid on each plot in yellow are the observed H12 and H2/H1 values for the top 25 peaks for each scan. We tested the effect of varying recombination rates on the Bayes factor analysis: (A)  $\rho = 0.5$  cM/Mb, and in (B)  $\rho = 0.5$  cM/Mb. In (A) and (B) H12 was measured in an analysis window centered on the selected site. Panels (A) and (B) show that the H12 and H2/H1 values measured at the top 25 peaks in Zambian data are supported by both hard and soft sweeps, whereas the H12 and H2/H1 values measured in Raleigh are supported by soft sweeps than hard sweeps in all scenarios. Varying  $\rho$  changes a small number of peaks identified as hard versus soft sweeps in the Zambian data. These peaks have low H12 value to begin with, making them difficult to distinguish between hard and soft sweeps definitively. In (C) instead of measuring H12 in a window centered on the selected site, the analysis window with the highest H12 value was first identified in a simulation of 100Kb, and then H2/H1 was measured. Here,  $\rho = 0.5$  cM/Mb. In this scenario, a few peaks in the Raleigh data are borderline supported by soft sweeps, but the results in Zambia do not change.



**Figure 6.** H12 and H2/H1 measured in sweeps of varying selective strength. Hard and soft sweeps were simulated with  $\theta_A$  of 0.01 and 10, respectively, and with selection strengths ( $s$ ) ranging from  $10^{-6}$  to 1. All simulations were conducted with a  $\rho=0.5$  cM/Mb in a sample size of 145 individuals and with a constant  $N_e \approx 2.6 \times 10^6$  model fit to Raleigh data (see Methods). Each boxplot comprises 100 simulations with the frequencies of the adaptive allele ranging from 0.4 to 0.6. The yellow region shows the range of H12 and H2/H1 values observed in Raleigh data at the top 25 peaks. The red line shows the one-per-genome false discovery rate value of H12 measured under neutrality with the  $N_e \approx 2.6 \times 10^6$  model (Text S1). H12 values are distinguishable from neutrality only when  $s \geq 10^{-3}$ . Hard sweeps can generate high H12 and H2/H1 values in the scenario when  $s=10^{-3}$ . However, this changes when  $\rho=5$  cM/Mb (Figure S8), whereby sweeps are distinguishable from neutrality only when  $s \geq 10^{-2}$  and hard sweeps can generate high H12 and some high H2/H1 values when  $s=10^{-2}$ .

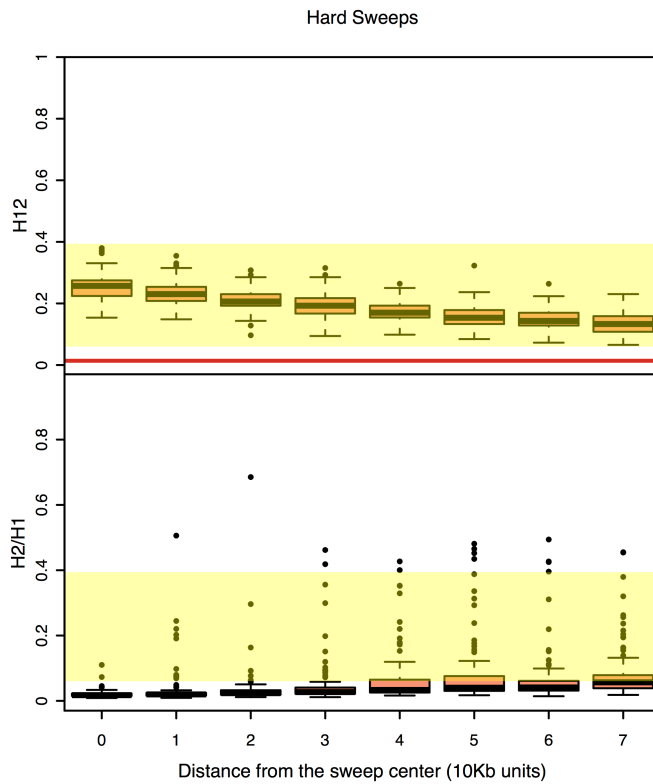


**Figure 7.** H12 and H2/H1 measured in sweeps decayed for varying amounts of time. Partial hard and soft sweeps were simulated with  $s=10^{-1}$  and  $\rho=0.5$  cM/Mb. All sweeps ended at a frequency of 0.5 when selection ceased. Sweeps were allowed to decay for varying amounts of time after selection ceased ( $T_E$ ), ranging from 0 to  $0.1 \cdot 4N_e$  generations. Simulations were performed in a sample size of 145 individuals and with a constant  $N_e \approx 2.6 \times 10^6$  model fit to Raleigh data (see Methods). Each boxplot comprises of 100 simulations. The yellow region shows the range of H12 and H2/H1 values observed in Raleigh data at the top 25 peaks. The red line shows the one-per-genome false discovery rate value of H12 measured under neutrality with the  $N_e \approx 2.6 \times 10^6$  model. In most evolutionary scenarios, hard sweeps cannot produce both high H12 and H2/H1 values, unlike soft sweeps. Hard sweeps that have not decayed for more than  $0.005 \cdot 4N_e$  generations have H12 values that are large enough to be differentiated from neutrality. Hard sweeps that have decayed for  $0.005 \cdot 4N_e$  generations have high H2/H1 values, but the H12 values in this scenario are borderline distinguishable from neutrality. Figure S8 shows additional evolutionary scenarios.





**Figure 8.** H12 and H2/H1 measured in hard sweeps at varying distances from the selected site. Partial hard sweeps were simulated with  $s=10^{-2}$  and  $\rho=0.5$  cM/Mb. All sweeps were sampled when selection ceased and had a frequency of 0.5. Simulations were performed in a sample size of 145 individuals and with a constant  $N_e \approx 2.6 \times 10^6$  model fit to Raleigh data (see Methods). Each boxplot comprises 100 simulations with H12 and H2/H1 values measured in 401 SNP windows centered varying distances away from the selected site. The yellow region shows the range of H12 and H2/H1 values observed in Raleigh data at the top 25 peaks. The red line shows the one-per-genome false discovery rate value of H12 measured under neutrality with the  $N_e \approx 2.6 \times 10^6$  model. The center of the sweep whether the selected site is located generates the highest H12 values. Analysis windows in the shoulders of the hard sweeps further away from the sweep center generate higher H2/H1 values. However, conditioning on the largest H12 value in a peak (defined as a consecutive tract of analysis windows with H12 values greater than the one-per-genome FDR values) results in low H2/H1 values for hard sweeps. Figure S8 shows similar results for complete sweeps and partial sweeps that have decayed for some time.



## TABLES

**Table 1:  $S$  and  $\pi$  measured in neutral demographic models of Zambian and Raleigh populations of *D. melanogaster*.** (A)  $S$  and  $\pi$  estimates for Zambian short intron data and demographic models fit to Zambian short intron data. (B)  $S$  and  $\pi$  estimates for Raleigh short intron data and demographic models fit to Raleigh short intron data. Estimates of  $S$  and  $\pi$  were averaged over 30,000 simulations of 10,000 bps for each demographic model.

### (A) Zambia

<b>Data/model</b>	<b><math>S</math>/bp</b>	<b><math>\pi</math>/bp</b>
ZI short intron data	10.1%	1.6%
Constant $N_e=4.8 \times 10^6$ model	9.9%	1.9%
Growth model	10.2%	1.1%
Bottlegrowth model	12.1%	1.7%

### (B) Raleigh

<b>Data/model</b>	<b><math>S</math>/bp</b>	<b><math>\pi</math>/bp</b>
RA short intron data	5.8%	1.2%
Admixture model	5.8%	1.1%
Admixture + bottleneck model	5.6%	1.3%
Constant $N_e=10^6$ model	2.3%	0.4%
Constant $N_e=2.6 \times 10^6$ model	5.8%	1.1%
Severe short bottleneck model	5.7%	1.1%
Shallow long bottleneck model	5.5%	1.1%

## SUPPLEMENT

### Text S1. Additional Materials and Methods

#### *Quality filtering of the Raleigh and Zambian data*

The Raleigh strains used for the analysis were generated by MACKAY *et al.* 2012 and belong to the *Drosophila* Genetic Reference Panel. The DGRP data set consists of fully sequenced genomes of 205 inbred *D. melanogaster* lines collected from Raleigh, North Carolina. The Zambian strains belong to the *Drosophila* Genome Nexus (LACK *et al.* 2015) and consist of 197 fully sequenced strains.

We masked regions in the data inferred to have high levels of identify by descent (IBD), admixture, and inversions because these are all factors that can contribute to high LD. Scripts for masking were provided by John Pool ([www.johnpool.net](http://www.johnpool.net)) along with regions inferred to have high IBD and admixture (LACK *et al.* 2015). Only African strains were masked for admixture between African and European strains. North American strains have a high level of genome-wide admixture (DUCHEN *et al.* 2013), and so, admixture in these strains was controlled for by performing neutral simulations under the admixture model by DUCHEN *et al.* 2013. All heterozygous sites and inferred tracts of residual heterozygosity (LACK *et al.* 2015) were masked as well. Inversions were masked by excluding the intervals of breakpoints identified in CORBETT-DETIG AND HARTL 2012; HUANG *et al.* 2014; LACK *et al.* 2015.

The extensive masking resulted in large tracts of ‘Ns’ in the data with an average length of 4 MB. In some cases, entire chromosomes were masked. However, strains with high levels of masking on one chromosome did not necessarily have high levels of masking on other chromosomes. To account for this heterogeneity in the strains with masking on each chromosome, we down sampled to 100 strains with the least number of Ns per chromosome. Thus, each chromosome consisted of different sets of 100 strains. The strain IDs used for each chromosome are given in Table S3.

To have the best resolution and power to detect individual selective sweeps with H12, we applied less stringent filters to the Raleigh and Zambian data sets and did not mask regions of IBD, admixture, and inversions. Instead, we excluded eight Zambian and 27 Raleigh strains with genome-wide IBD levels with at least one other strain  $> 20\%$  as calculated by LACK *et al.* 2015. The excluded strain IDs are as follows: ZI397, ZI530, ZI269, ZI240, ZI218, ZI207, ZI523, ZI86,

RAL-385, RAL-358, RAL-712, RAL-399, RAL-879, RAL-355, RAL-810, RAL-350, RAL-832, RAL-882, RAL-306, RAL-799, RAL-801, RAL-859, RAL-907, RAL-790, RAL-748, RAL-336, RAL-850, RAL-365, RAL-786, RAL-730, RAL-861, RAL-59, RAL-646, RAL-812, and RAL-787. In this version of the data set, we masked all heterozygous sites and inferred tracts of residual heterozygosity as in the previous data set.

H12 was applied to the entire 178 Raleigh strains and 188 Zambian strains. To reduce any remaining heterogeneity in sample size due to missing data, we down-sampled to the 145 strains with the least amount of missing data in each analysis window for inclusion in the H12 calculation.

### ***Calculation of the critical value (H12<sub>c</sub>)***

To calculate the critical value of H12, or, 1-per-genome false discovery rate value, we used the method described in GARUD *et al.* 2015 whereby we simulated 130,000 analysis windows under a neutral constant  $N_e$  demographic model, which represents approximately 10x the number of analysis windows in the data. We assigned the 10<sup>th</sup> highest H12 value in the simulated distribution to be H12<sub>c</sub>.

### ***Calculation of the probability of overlap of peaks in the Zambian and Raleigh scans***

To calculate the probability of 3 or greater overlapping peaks in the Zambian and Raleigh scans in Figure 2, we calculated the fraction of the autosomal arms in each genome covered by H12 peaks. In Raleigh 1.51% of the autosomal arms are covered by H12 peaks (corresponding to 4,252,692 bps) and in Zambia 4.38% of the autosomal arms are covered by H12 peaks (corresponding to 1,461,431bps). The probabilities of observing three out of 25 peaks in Raleigh overlapping three out of 25 peaks in Zambia can be computed with a right-tailed binomial probability, where the probability of an overlap in Zambian or Raleigh data corresponds to either 1.51% or 4.38%, respectively. Thus the probability of three or more overlaps in the Raleigh data is 0.0224 and the probability of 3 or more overlaps in the Zambian data is 0.0005.

### ***Test for structure, admixture and backflow generating H12 peaks in Zambian data***

We tested for an enrichment of strains among the top three haplotypes across all the peaks in the Zambian H12 scan. To calculate significance, we performed a permutation analysis. We shuffled

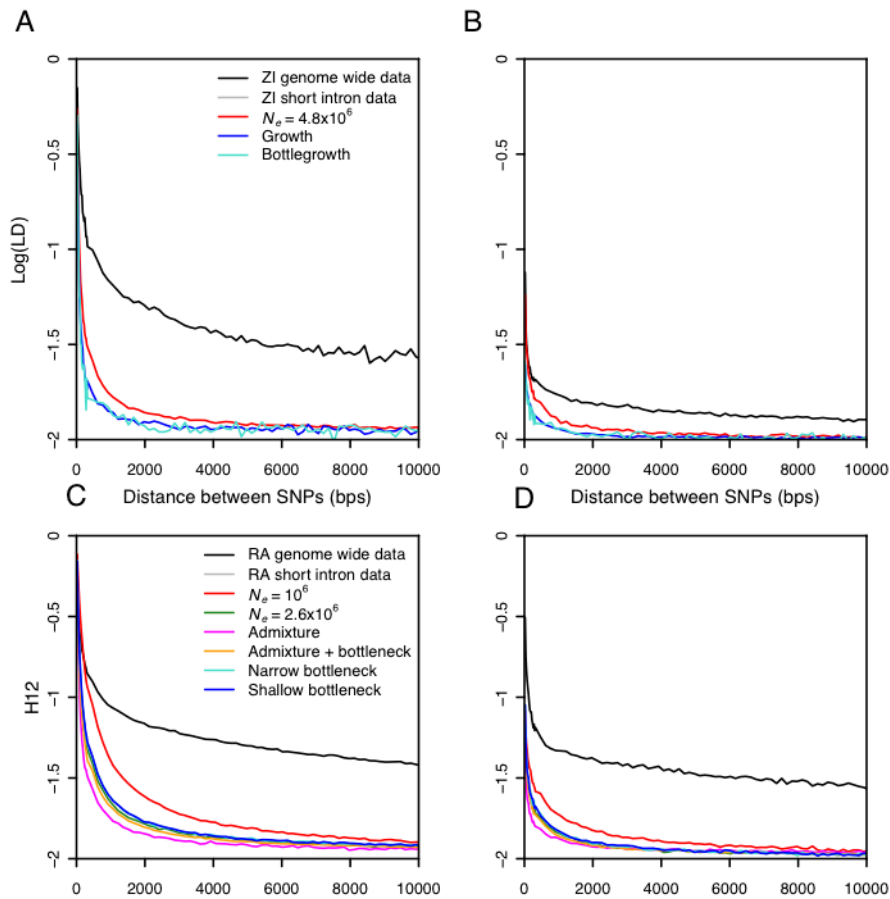
the labels of strains comprising all the haplotypes at each peak and counted the number of strains appearing among the three most common haplotypes in  $n$  or more peaks, where  $n$  ranged from 0 to 25 peaks. We repeated this procedure 10,000 times. Then, comparing against the distribution of the number of repeated strains among the top three haplotypes for  $n$  or more peaks, we calculated an empirical P-value for observing 15 strains appearing among nine or more peaks' top haplotypes. We did the same for all other  $ns$  and found that for lower  $ns$  we did not get significant P-values.

We tested for the possibility of backflow of strains from North America to Zambia generating the H12 peaks in the Zambian data set. First we tested for a dip in diversity in the Raleigh data at positions corresponding to the top 25 peaks in the Zambian data set. We computed  $\pi$  in 10Kb non-overlapping windows on all autosomal arms in the Raleigh data. We then calculated the mean  $\pi$  in any window overlapping the start and end coordinates of each of the top 25 peaks identified in the Zambian data set and compared this mean with the distribution of  $\pi$  calculated in all remaining windows in the Raleigh data set.

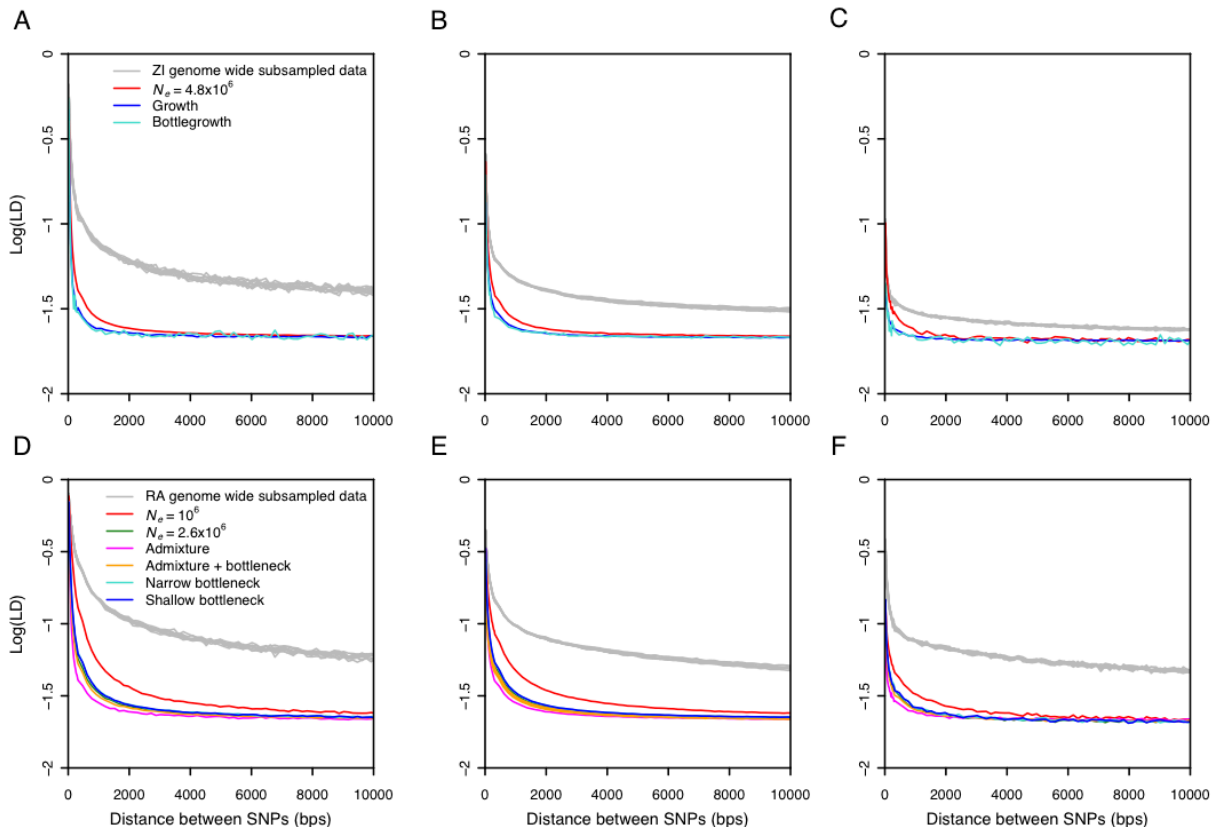
### ***Test for enrichment of peaks in ancestry disequilibrium hubs***

To test whether the top 25 peaks in the Raleigh data set are enriched in ancestry disequilibrium hubs identified by POOL 2015, we first identified that the end coordinates of the first and last windows of two peaks overlap the coordinates of the hubs listed in Table S6 of POOL 2015. We then calculated the fraction of the autosomes that the hubs occupy (8.1%) and then calculated a binomial probability of observing two out of 25 peaks overlapping any hub at random.

**Figure S1.** Elevation in LD values in Raleigh and Zambian data relative to neutral demographic models. In these panels, minor allele frequencies for SNPs were conditioned on values between 0.3 and 0.5, and 0.0 and 0.05. (A) Zambian (ZI) data and demographic models with MAFs between 0.3 and 0.5. (B) Zambian data and demographic models with MAFs between 0.0 and 0.05. (C) Raleigh (RA) data and demographic models with MAFs between 0.3 and 0.5. (D) Raleigh data and demographic models with MAFs between 0.0 and 0.05.

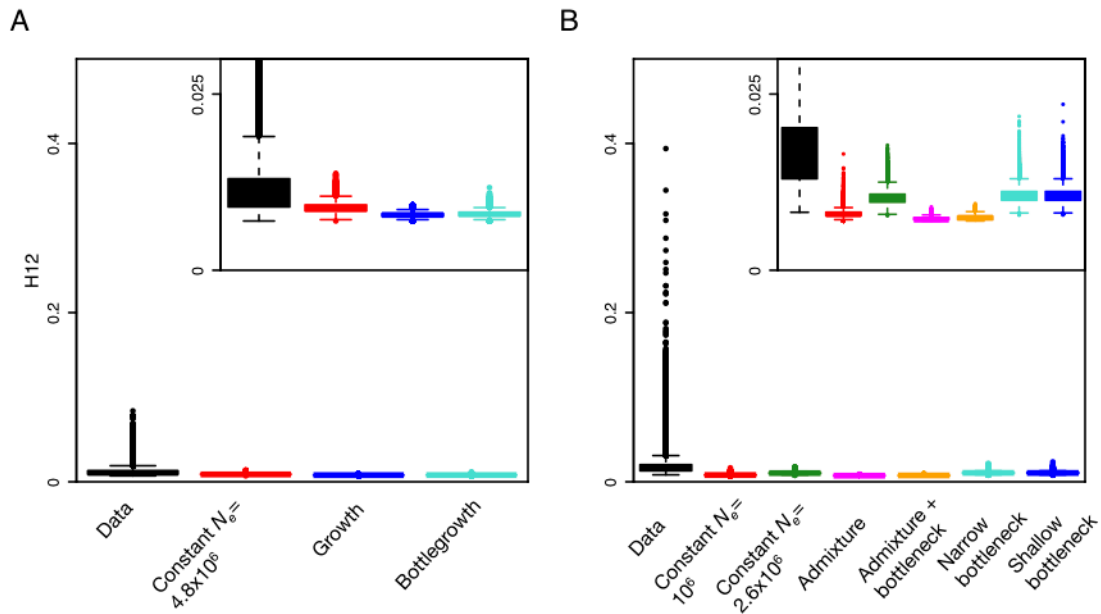


**Figure S2.** Elevation in LD values in Zambian and Raleigh data after down sampling to 50 strains. Zambian and Raleigh strains were down-sampled to 50 strains 10 times and the resulting LD distributions were plotted (grey). In contrast to expectations under any neutral demographic model tested with a sample size of 50, all samples of 50 strains have elevated long range LD. This indicates that the elevation in LD is not driven by any sub-population. The LD calculations were conditioned on minor allele frequency (MAF) classes as follows: (A) Zambian (ZI) data and demographic models with MAFs between 0.3 and 0.5. (B) Zambian data and demographic models with MAFs between 0.05 and 0.5. (C) Zambian data and demographic models with MAFs between 0.0 and 0.05. (D) Raleigh (RA) data and demographic models with MAFs between 0.3 and 0.5. (E) Raleigh data and demographic models with MAFs between 0.05 and 0.5. (F) Raleigh data and demographic models with MAFs between 0.0 and 0.05.



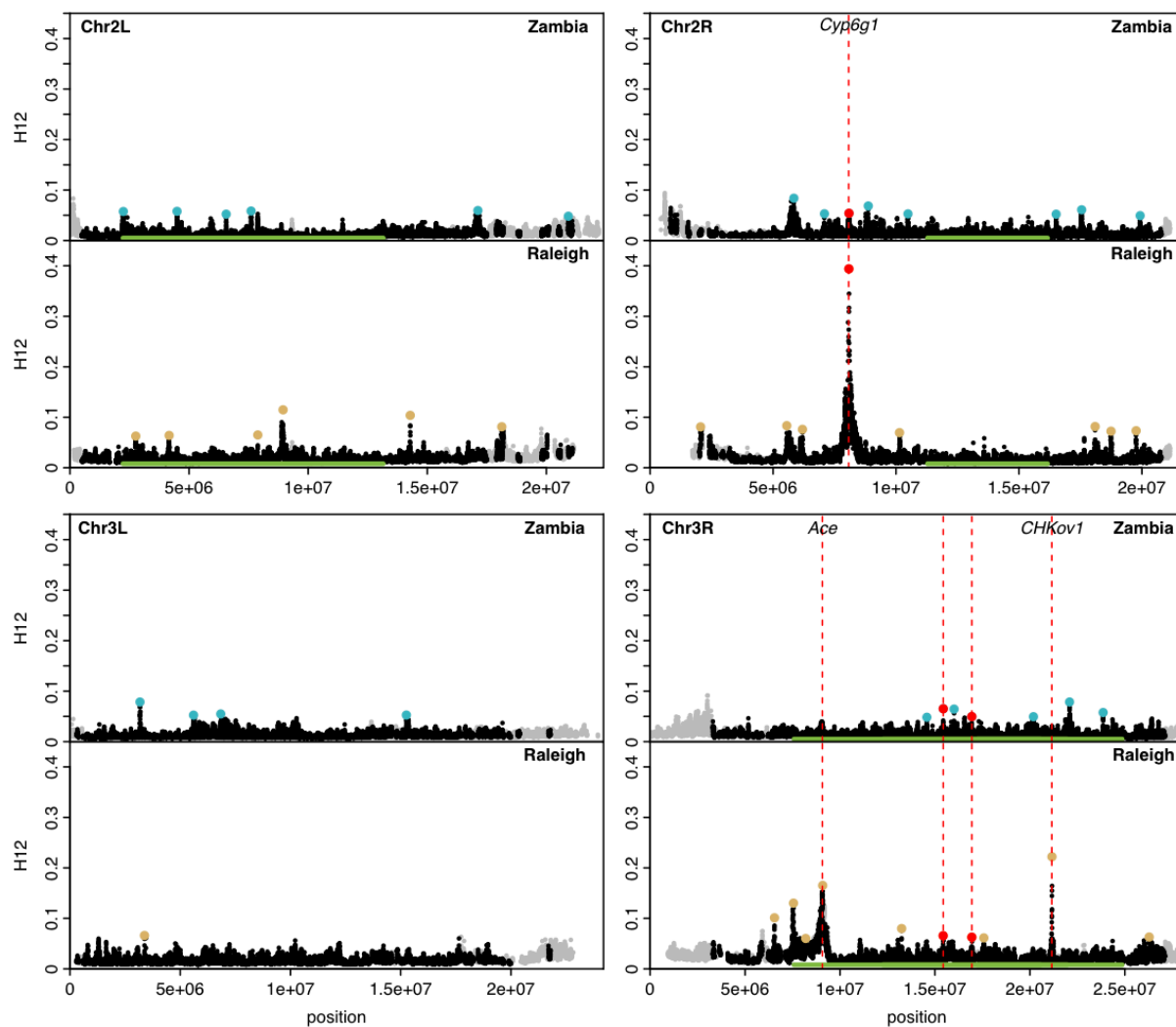
**Figure S3.** Elevation in H12 values in Zambian and Raleigh data with less stringent filters.

H12 values measured in windows of 801 SNPs down sampled to 401 SNPs in Zambian data (A) windows of 401 SNP in Raleigh data (B) along the four autosomal arms were elevated as compared to expectations under any neutral demographic model tested. In this data set, we did not mask inversions, IBD tracts, or admixture tracts. Instead, we excluded individuals with > 20% of the autosomal arms sharing IBD tracts of 0.5 similarity or greater with any other another individual in the sample. H12 values were measured in sample sizes of 145 and in genomic regions with  $\rho \geq 0.5$  cM/Mb. H12 values were measured in neutral demographic simulations of 145 individuals generated with  $\rho = 0.5$  cM/Mb. Plotted are the result of approximately  $1.5 \times 10^5$  simulations under each neutral demographic model, representing 10 times the number of analysis windows in the data.

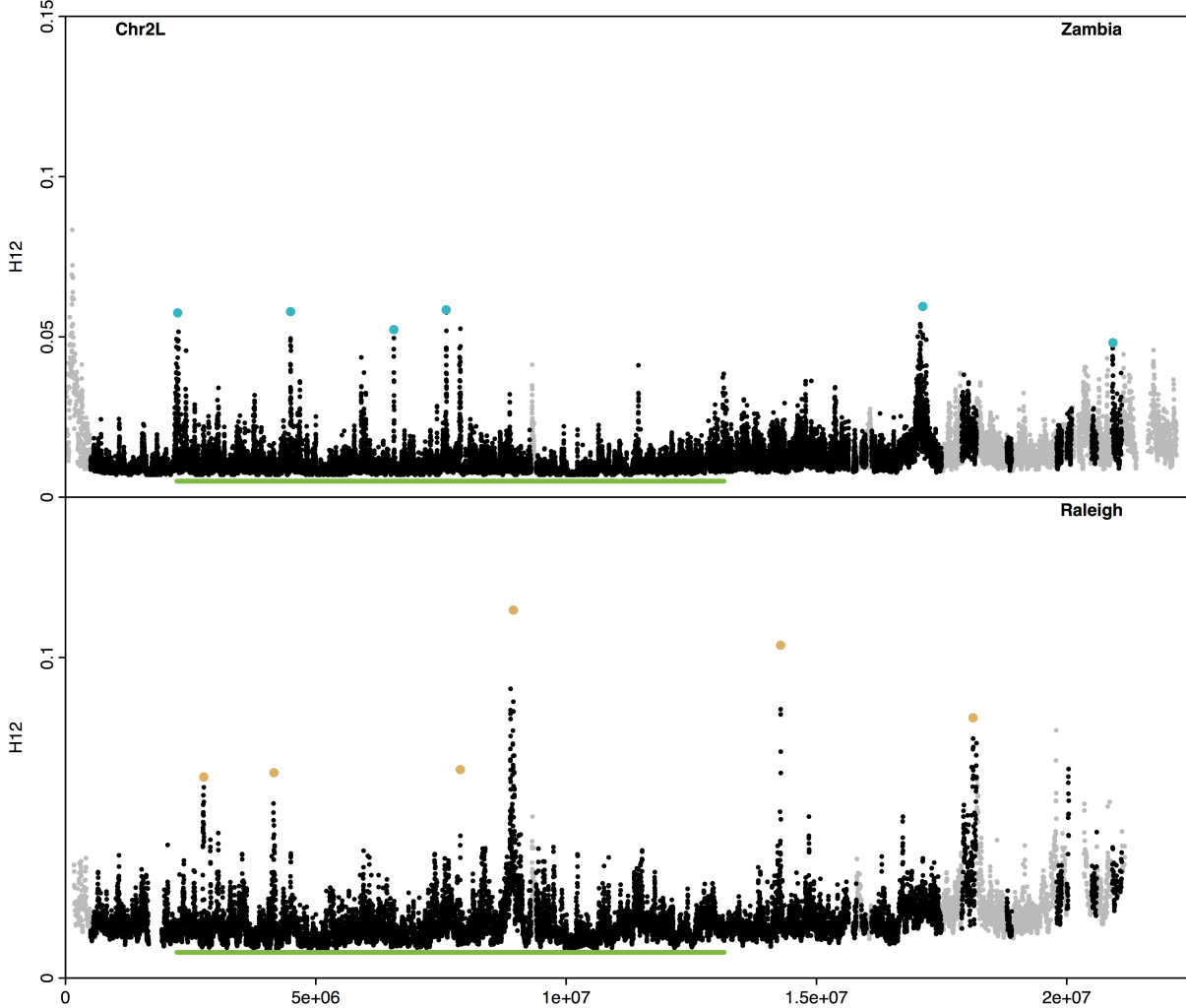


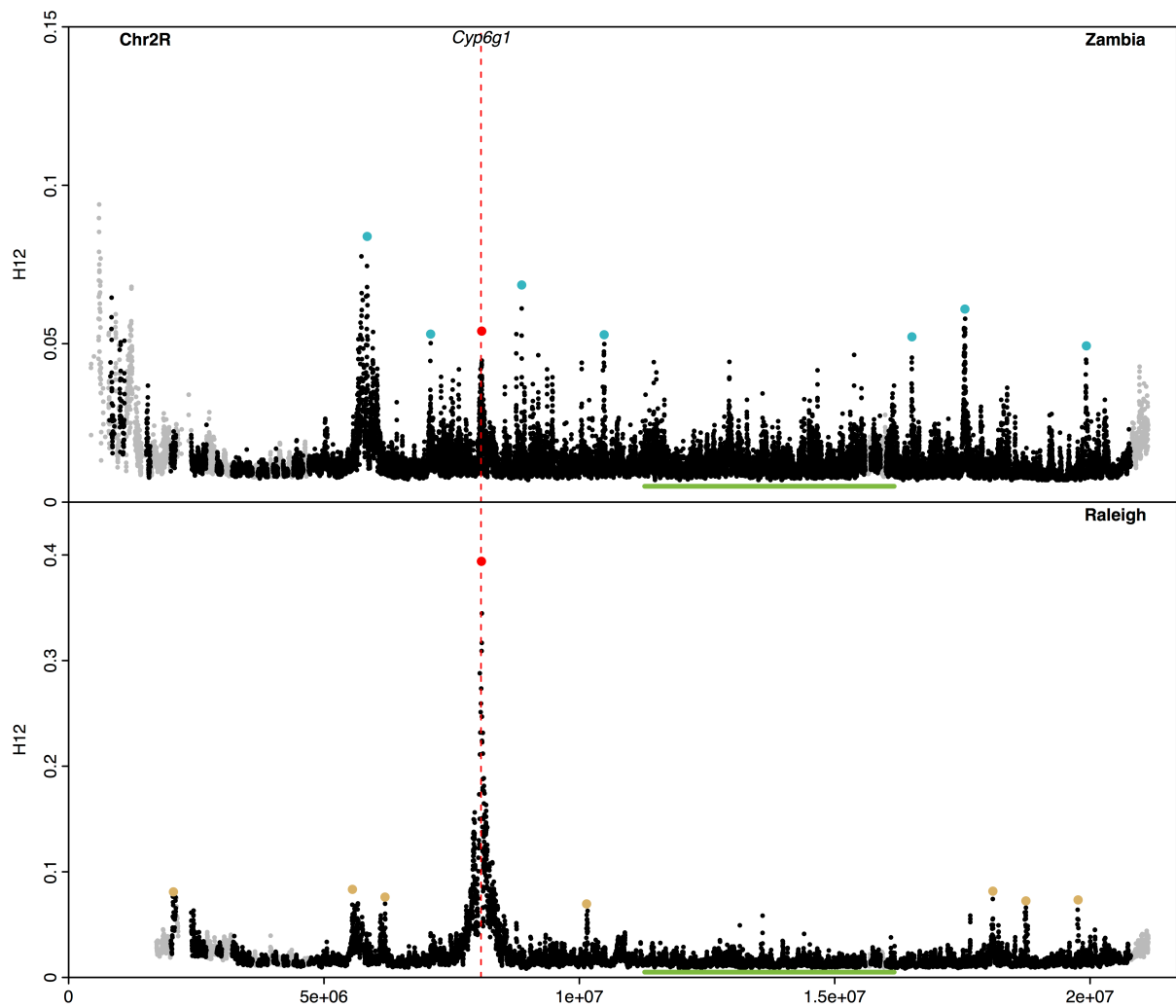


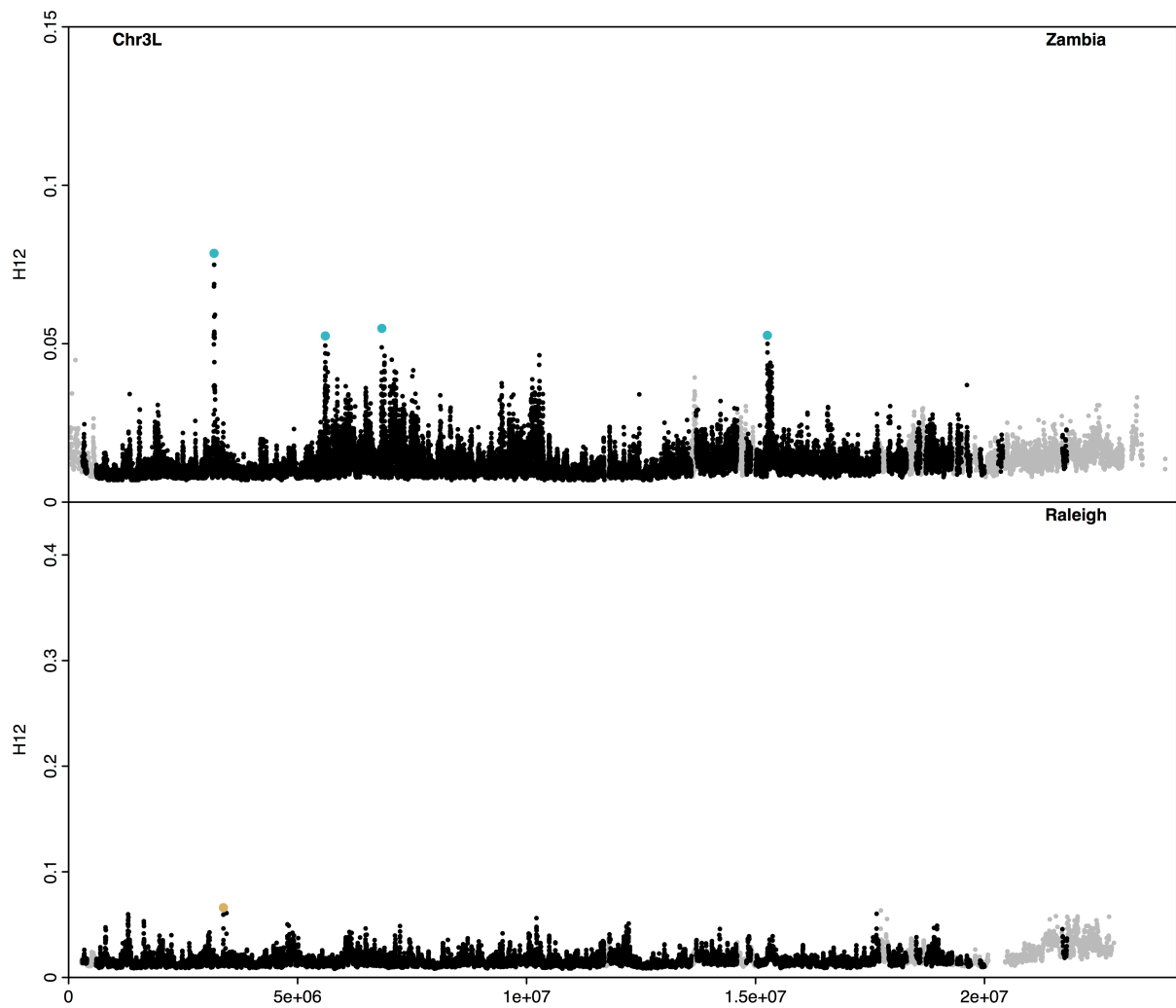
**Figure S4.** H12 scan in Zambian and Raleigh data. This figure is the same as Figure 2, except here the range of the y-axis is the same in both scans.

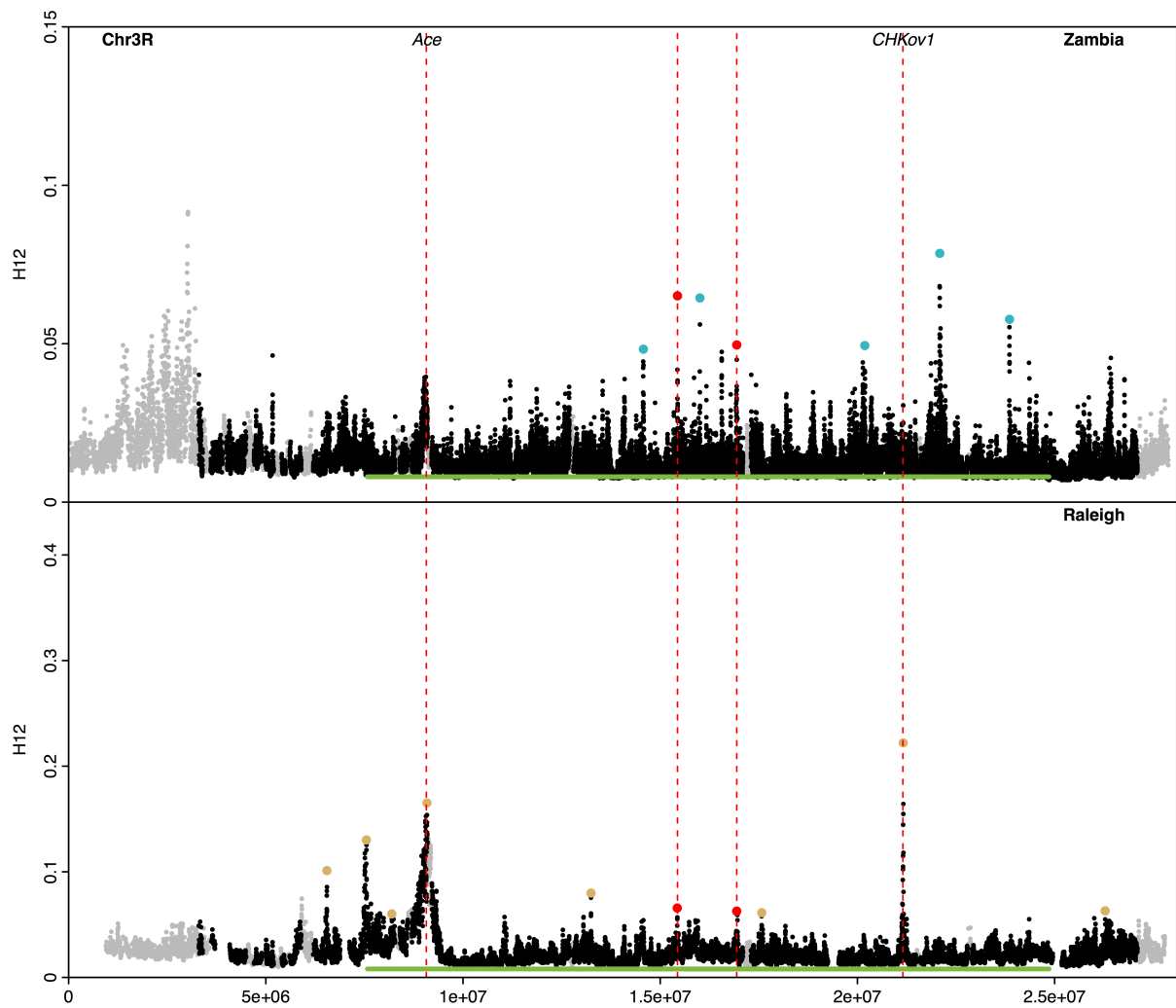


**Figure S5.** H12 scan in Zambian and Raleigh data zoomed in for higher resolution. This figure is the same as Figure 2.

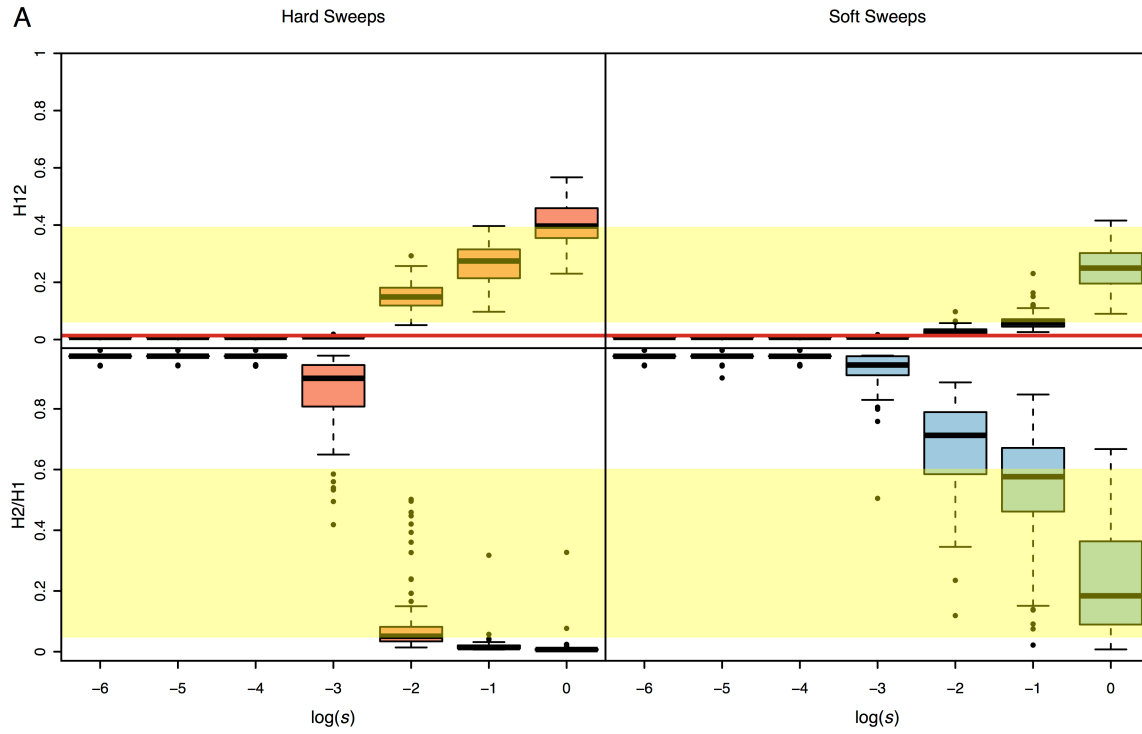




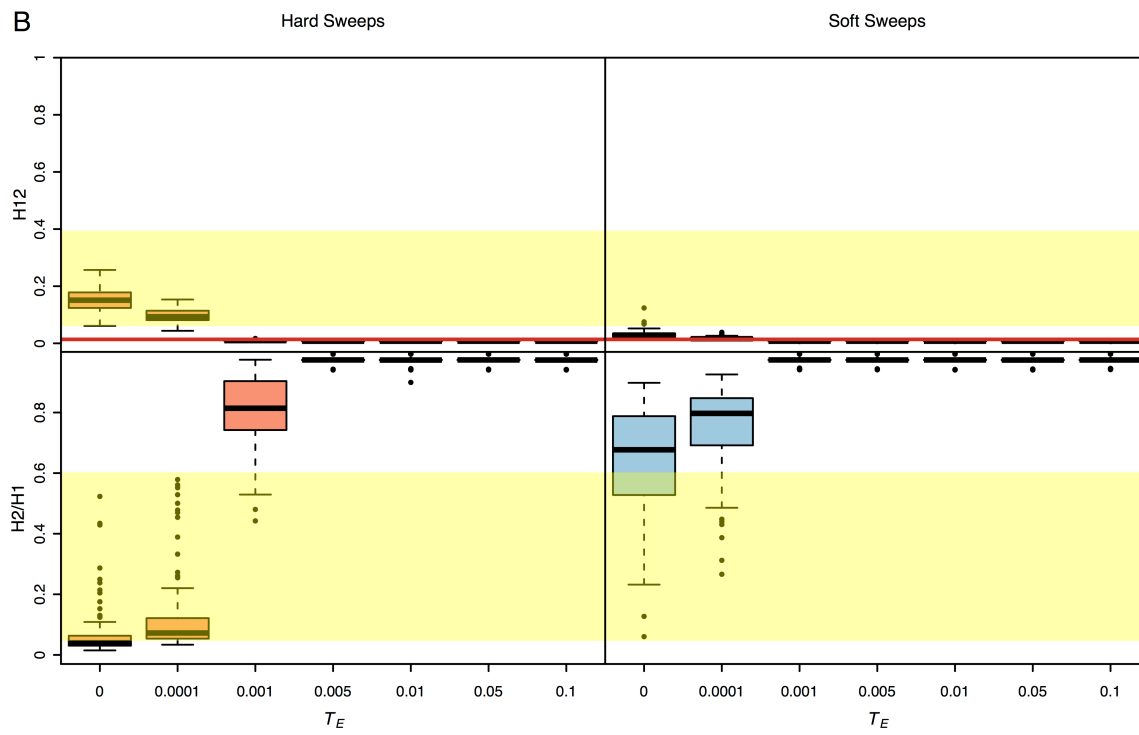
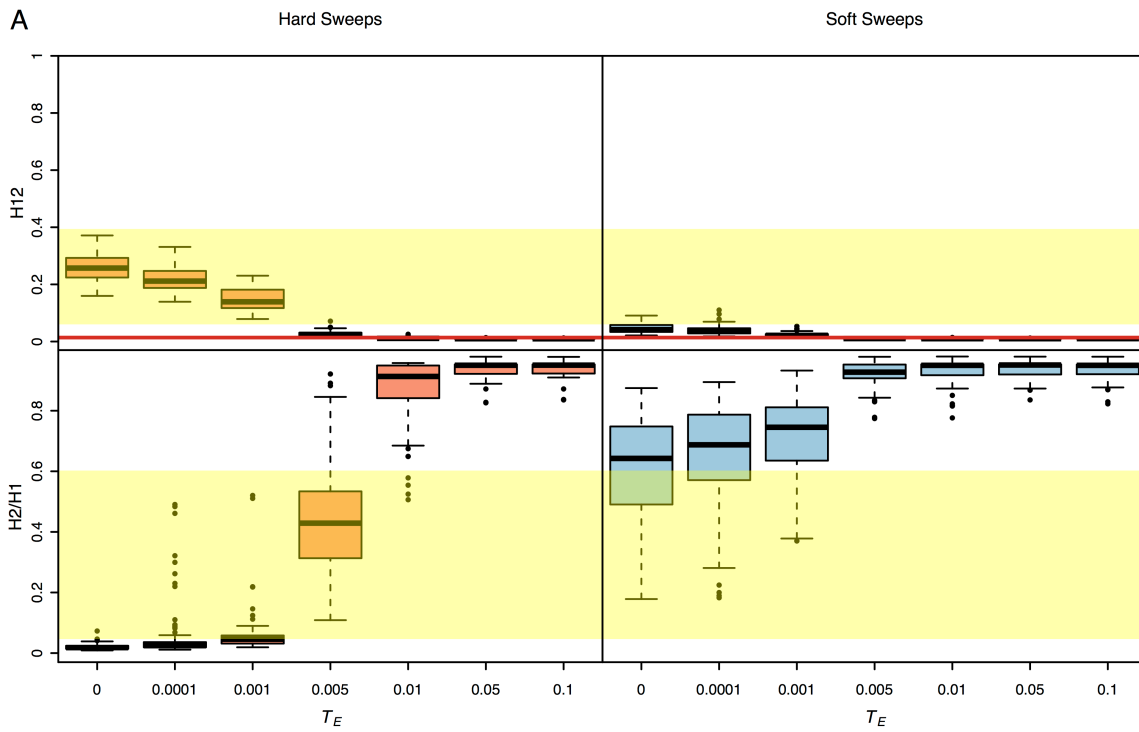




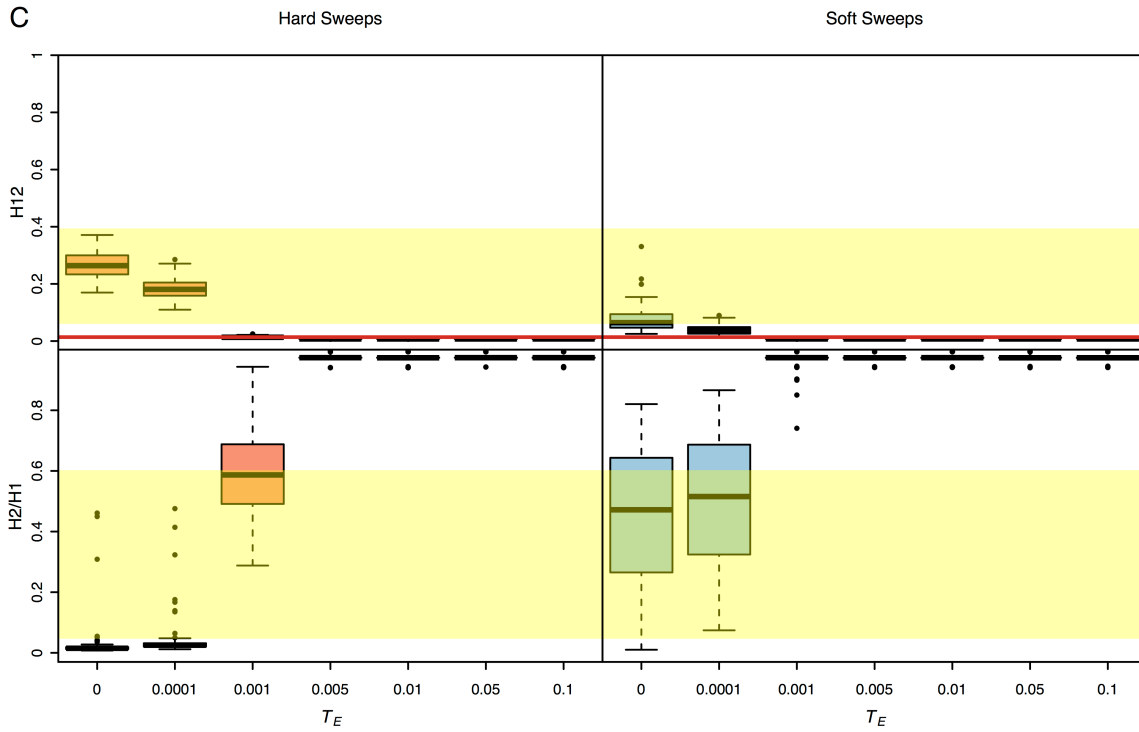
**Figure S6.** H12 and H2/H1 measured in sweeps of varying selective strength. Simulations were performed with the same parameters as in Figure 6 except with a higher recombination rate of  $\rho=5$  cM/Mb.



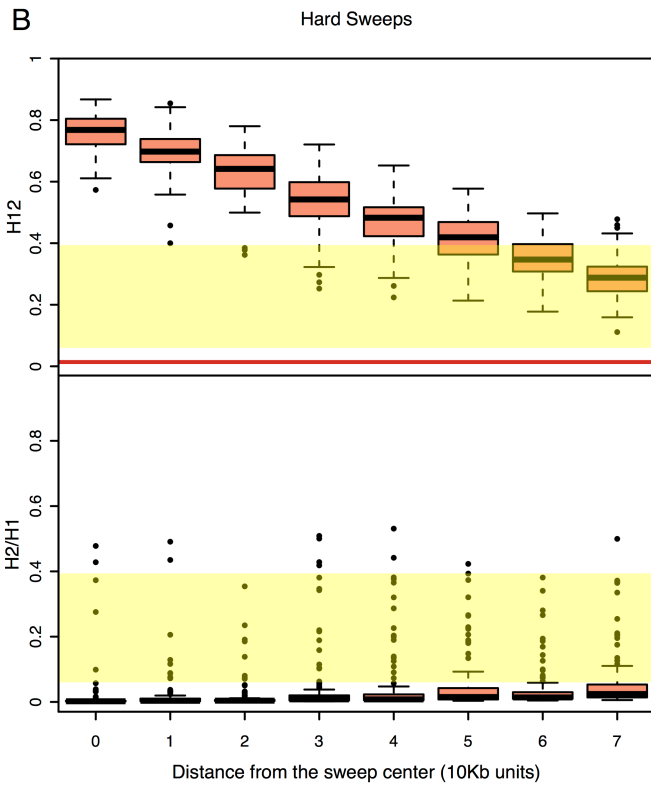
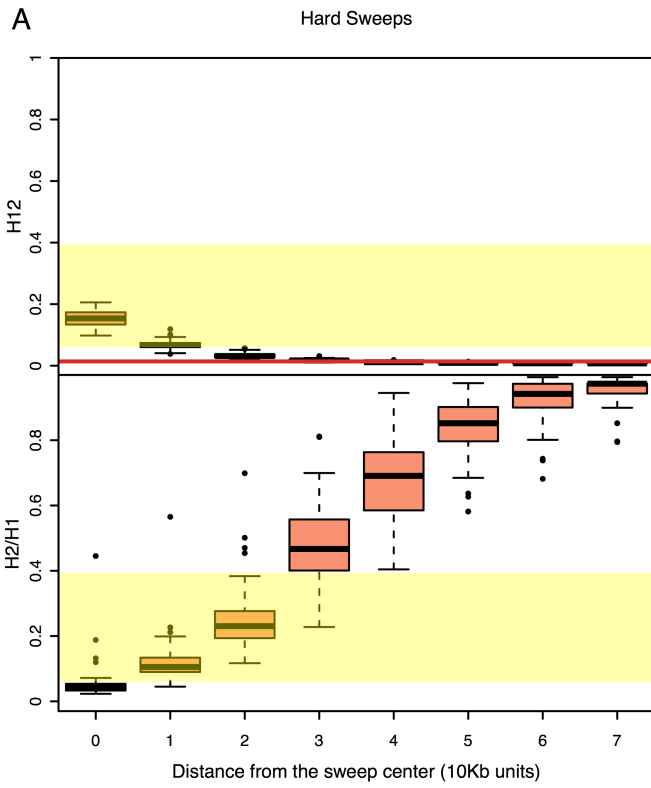
**Figure S7.** H12 and H2/H1 measured in sweeps decayed for varying amounts of time in additional evolutionary scenarios. Simulations were performed with the same parameters as in Figure 7 with some modifications. In (A)  $s=0.01$  and  $\rho=0.5$  cM/Mb. In (B)  $s=0.01$  and  $\rho=5$  cM/Mb. In (C)  $s=0.1$  and  $\rho=5$  cM/Mb.







**Figure S8.** H12 and H2/H1 measured in hard sweeps at varying distances from the selected site in additional evolutionary scenarios. Simulations were performed with the same parameters as in Figure 8 with some modifications. In (A) the sweeps were decayed for  $0.001 * 4N_e$  generations and the frequency of the adaptive mutation was 0.5 at the time when selection ceased. In (B) sweeps are complete and were sampled when selection ceased without any time for decay. In all scenarios tested, the H12 value at the center of the sweep where the selected site is located is the highest and H2/H1 is the lowest.



**Table S1. Demographic parameters inferred by DaDi for models fit to Zambian data.**

Shown are parameter estimates and corresponding log likelihoods (LL) for the bottleneck, bottlegrowth, and growth models inferred by DaDi using Zambian short intron data

GUTENKUNST *et al.* 2009.  $N_B$  represents the bottleneck size,  $N_F$  the final population size,  $T_B$  the duration of the bottleneck, and  $T_F$  the duration of the final population size. For the growth model, both  $N_F$  represents the final population size and  $T$  the time at which growth started. All population size estimates are in terms of units  $4 * N_{e_{ancestral}}$  and all time estimates are in terms of units  $2 * N_{e_{ancestral}}$ .

**Bottleneck model:**

$N_B$	$N_F$	$T_B$	$T_F$	LL
1.402	2.940	0.203	0.048	-214.476

**Bottlegrowth model:**

$N_B$	$N_F$	$T_B$	$T_F$	LL
1.465	3.388	0.128	0.116	-214.376

**Growth model:**

$N_F$	$T$	LL
2.87	0.17	-228.54

**Table S2. IDs of strains included in the analysis of LD and haplotype homozygosity patterns in the stringently filtered Zambian and Raleigh data sets.**

We down-sampled the Zambian and Raleigh strains used for the analysis of LD and haplotype homozygosity in Figures 1 and S1 to the 100 strains with the least amount of missing data for each population. This down sampling was done after extensive masking of regions of IBD, admixture, and inversions. Since the amount of missing data varied from chromosome to chromosome depending mainly on the presence of inversions, we down-sampled each chromosome separately. The columns in this table show the strains that were included for the analysis for each chromosome for each population.

<b>Table S3. Test for elevation in LD within 100 Kb of inversion breakpoints.</b>		
<b>Population</b>	<b>Mean LD of SNP pairs within 100 Kb of inversion breakpoints</b>	<b>Mean LD of SNP pairs beyond 100 Kb of inversion breakpoints</b>
Zambia	0.0253	0.0268
Raleigh	0.0610	0.0602

Shown are mean LD values calculated for 100,000 randomly sampled SNP pairs within and beyond 100 Kb of inversion breakpoints, which were identified in CORBETT-DETIG AND HARTL 2012; HUANG *et al.* 2014; LACK *et al.* 2015. The differences in mean values are biologically insignificant.

**Table S4. Top 25 peaks in the Zambian (A) and Raleigh (B) data.** Listed are the coordinates of the center of the analysis window with the highest H12 value in a peak, the edge coordinates of each peak, and the corresponding H12 and H2/H1 values in the analysis window.

**(A) Zambia**

Chromosome	Start of peak	End of peak	Center of peak	H12	H2/H1
Chr2R	5689984	5877052	5845467	0.084	0.448
Chr3R	22074333	22099126	22087580	0.079	0.106
Chr3L	3143696	3207496	3175369	0.079	0.106
Chr2R	8858953	8877924	8871146	0.069	0.210
Chr3R	15430749	15451721	15437579	0.065	0.591
Chr3R	15997772	16011492	16006285	0.064	0.132
Chr2R	17495878	17569226	17549291	0.061	0.151
Chr2L	16962865	17287118	17124919	0.060	0.368
Chr2L	7595364	7622463	7603947	0.058	0.156
Chr2L	4481025	4519970	4495822	0.058	0.208
Chr3R	23843073	23866972	23856074	0.058	0.143
Chr2L	2224927	2247036	2241538	0.058	0.125
Chr3L	6820734	6858547	6839441	0.055	0.141
Chr2R	8072574	8129240	8087869	0.054	0.345
Chr2R	7075343	7105539	7088009	0.053	0.407
Chr2R	10472210	10508206	10485863	0.053	0.449
Chr3L	15243946	15273825	15258107	0.053	0.592
Chr3L	5576197	5623343	5604719	0.052	0.222
Chr2L	6551349	6571793	6558631	0.052	0.194
Chr2R	16497602	16522034	16512033	0.052	0.274
Chr3R	16920082	16949847	16940157	0.050	0.266
Chr3R	20157200	20203615	20186573	0.049	0.262
Chr2R	19900056	19941093	19927955	0.049	0.167
Chr3R	14551505	14588594	14572745	0.048	0.622

Chr2L	20855660	21041236	20923022	0.048	0.671
-------	----------	----------	----------	-------	-------

**(B) Raleigh**

Chromosome	Start of peak	End of peak	Center of peak	H12	H2/H1
Chr2R	7723122	8452931	8082481	0.394	0.046
Chr3R	21100160	21196327	21161195	0.222	0.388
Chr3R	8561176	9347112	9084900	0.165	0.338
Chr3R	7441836	7656600	7549332	0.130	0.098
Chr2L	8787750	8981351	8946244	0.115	0.524
Chr2L	14260585	14297414	14284521	0.104	0.330
Chr3R	6503887	6568950	6548287	0.101	0.149
Chr2R	5488736	5706118	5556716	0.083	0.558
Chr2R	18048793	18139982	18096795	0.082	0.501
Chr2L	18066440	18156804	18128033	0.081	0.229
Chr2R	1980165	2741000	2051987	0.081	0.516
Chr3R	13235515	13272233	13244854	0.080	0.580
Chr2R	6158540	6220781	6194741	0.076	0.524
Chr2R	19741265	19784611	19765222	0.073	0.179
Chr2R	18730636	18755472	18743856	0.072	0.326
Chr2R	10116253	10172867	10141807	0.070	0.602
Chr3L	3361643	3387951	3379415	0.066	0.305
Chr3R	15417427	15448395	15431116	0.066	0.555
Chr2L	7874465	7898299	7886142	0.065	0.386
Chr2L	4146117	4176261	4164595	0.064	0.409
Chr3R	26262373	26301949	26281108	0.063	0.566
Chr3R	16908455	16953298	16939738	0.063	0.245
Chr2L	2735006	2781702	2762777	0.063	0.313
Chr3R	17548069	17614825	17573640	0.061	0.480
Chr3R	8113141	8551014	8193336	0.060	0.413

**Table S5. Test for enrichment of peaks overlapping regions with inversions.** We tested for an enrichment of peaks in regions of the genome with inversions that are present in at least 10 strains in the Zambian and Raleigh data sets. To do so we performed a one-sided binomial test calculating the probability of at least the number of peaks observed to overlap assuming a uniform distribution of peaks along the autosomal arms. We observe an enrichment of peaks overlapping common inversions on chromosome 3R when performing the scan in windows of 801 SNPs down sampled to 401 SNPs in Zambian data and in windows of 401 SNPs in Raleigh data (low probabilities are highlighted in bold).

Inversion	Number of overlapping peaks in Zambia (scan performed in 801 SNPs windows down sampled to 401 SNPs)	Probability	Number of overlapping peaks in Raleigh (scan performed in 401 SNP windows)	Probability
In(2L)t	4	0.14	4	0.14
In(2R)ns	0	0.73	0	0.73
In(3L)Ok	3	0.20	-	-
In(3R)K	5	0.16	7	<b>0.02</b>
In(3R)Mo	-	-	2	0.31
In(3R)P	5	<b>0.02</b>	5	<b>0.02</b>