GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.116459/DC1

Searching for Footprints of Positive Selection in Whole-Genome SNP Data from Nonequilibrium Populations

Pavlos Pavlidis, Jeffrey D. Jensen and Wolfgang Stephan

DOI: 10.1534/genetics.110.116459

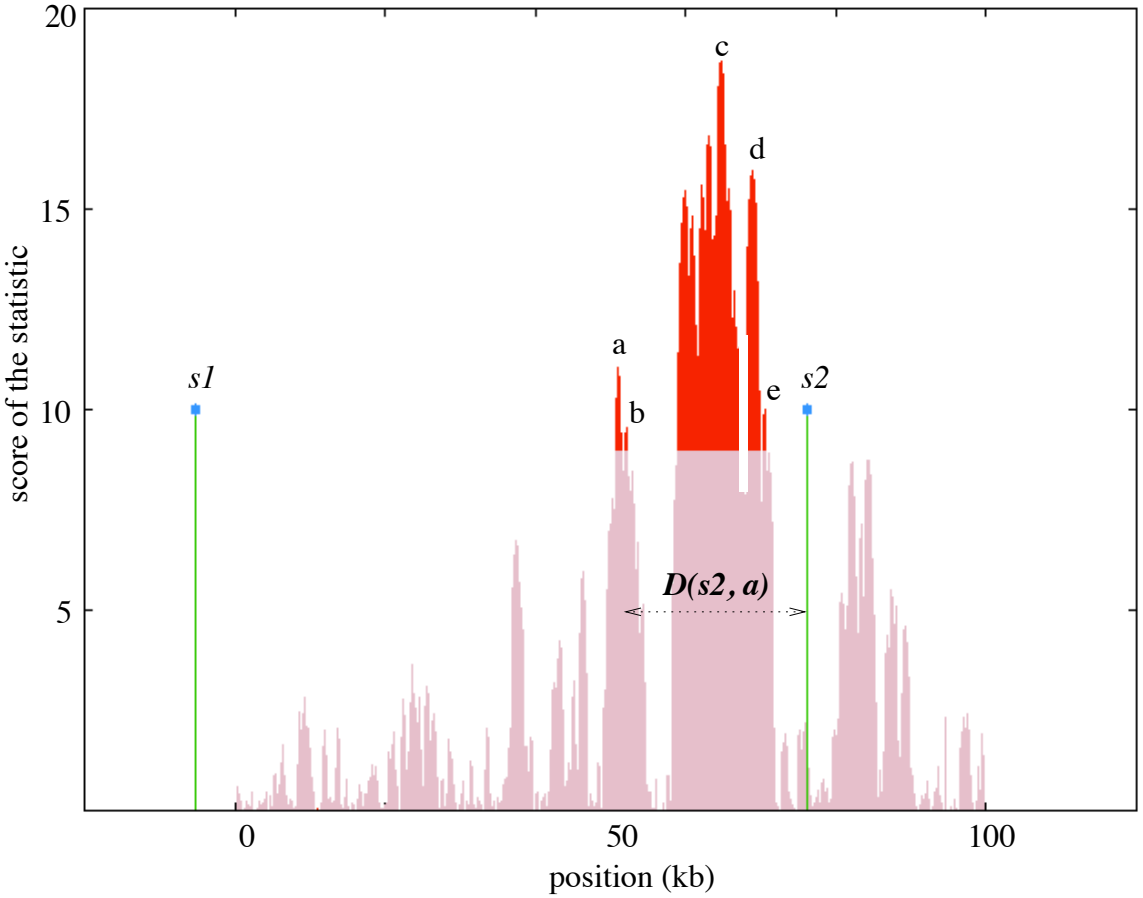

FIGURE S1.– The distance between a peak of the landscape of the statistic and the selective sweep locations. In the history of the population two selective sweeps have occurred recently, at different time points and different locations on the chromosome. The selective sweep locations are illustrated as *s1* and *s2* (vertical lines). Given a user defined threshold, the landscape of the statistic is split in two regions, *i.e.* above and below the threshold. A peak is defined as the maximum point in an isolated (by the threshold) region. Thus, 5 peaks (a to e) have been formed. The distance $D(s2, a)$ of the 'a' peak measures the distance between this peak and *s2* which is the closest sweep location from this peak.

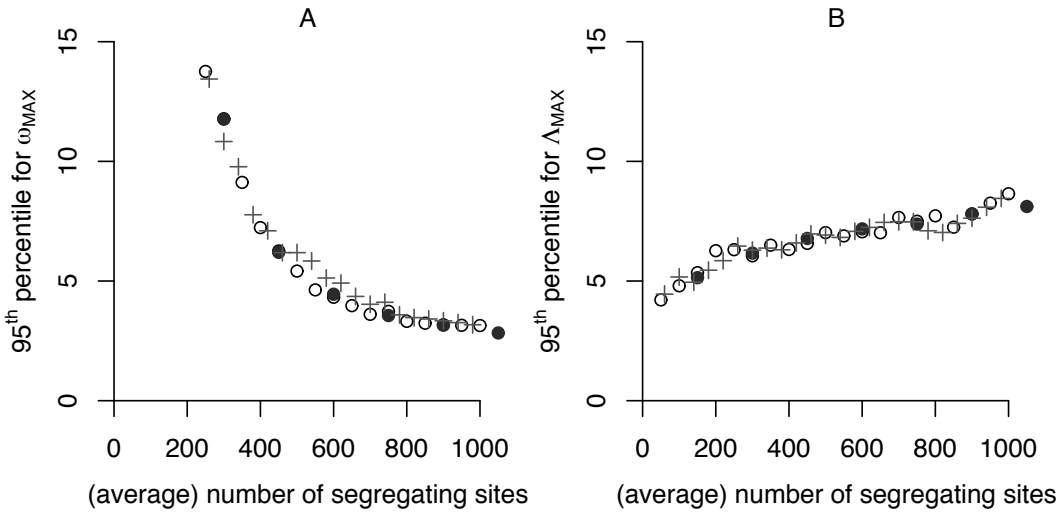

FIGURE S2.– The 95^{th} percentile for (A) the ω-statistic and (B) *SweepFinder* based on the $F\theta$ (filled circles), the FS (open circles) and the $F\theta S$ procedures (crosses). Equilibrium neutrality simulations have been performed for a 50-kb genomic segment and 12 sequences ($h_n \approx 3$). Recombination rate is 0.05/bp. For a given number of segregating sites (x-axis) simulations were performed by (i) fixing the number of segregating sites S_n (open circles), (ii) using $\theta_{\rm NEU} = \theta_{\rm W} = \frac{S_n}{h_n}$ (filled circles). In this case simulations generate on average S_n segregating sites. (iii) Under the $F\theta S$ process (crosses) we used the same $\theta_{\rm NEU} = \theta_{\rm W} = \frac{S_n}{h_n}$ but only the realizations that produced S_n segregating sites.

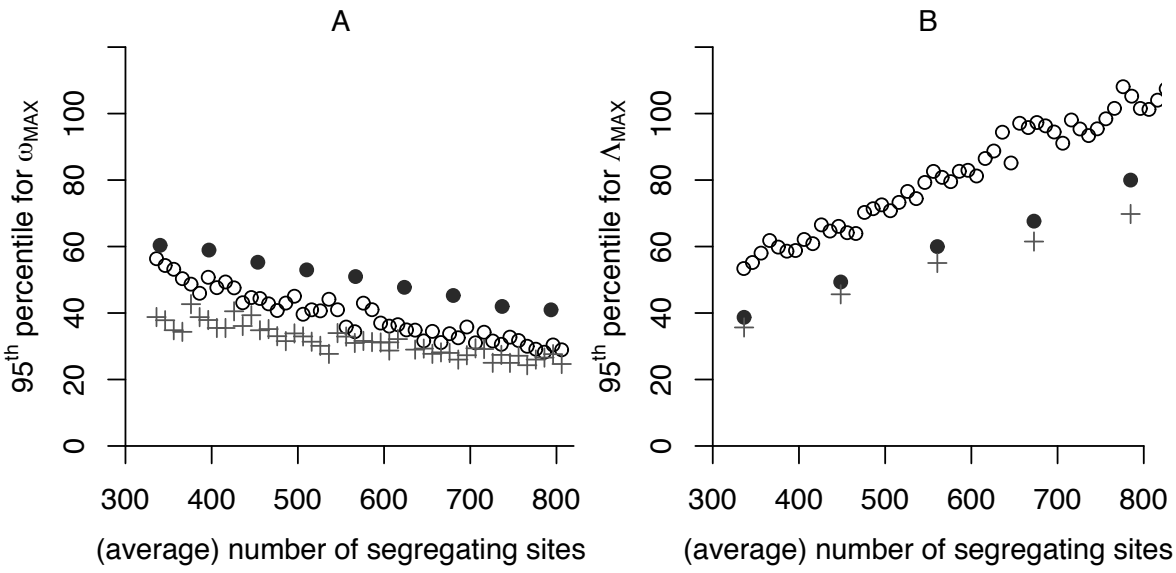

FIGURE S3.– The 95^{th} percentile for (A) the $\omega_{\rm MAX}$ and (B) $\Lambda_{\rm MAX}$ based on the $F\theta$ (full circles), the FS procedure (open circles) and the $F\theta S$ approach (crosses). Bottleneck simulations have been performed for a 50-kb genomic segment and 12 sequences ($h_n \approx 3$). We have used the demographic scenario inferred by LI and STEPHAN (2006) that describes the history of the European population of *D. melanogaster*. Recombination rate is 0.05/bp. For a given number of segregating sites (x-axis) simulations were performed by (i) fixing the number of segregating sites S_n (open circles), (ii) using $\theta_{\rm NEU} = \widehat{\theta} = \frac{2S_n}{E(T_c)}$, where $E(T_c)$ is the expected total length of the coalescent of n sequences (ZIVKOVIC and WIEHE, 2008) (filled circles). In this case simulations generate on average S_n segregating sites. (iii) Under the $F\theta S$ process (crosses) we used the same $\theta_{\rm NEU} = \widehat{\theta}$, but only the realizations that produced S_n segregating sites).

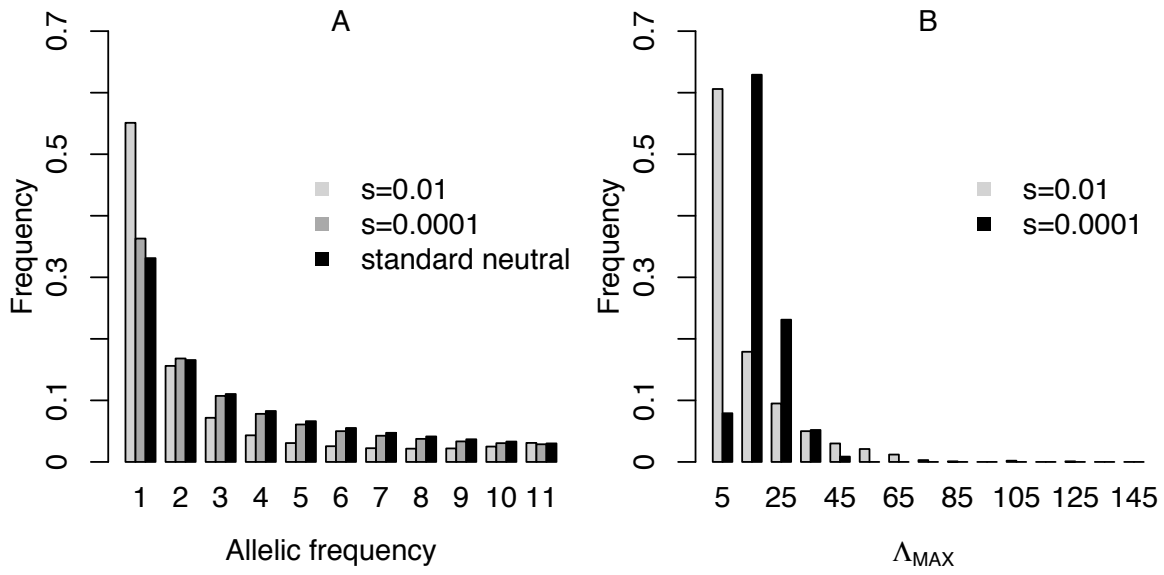

FIGURE S4.– Comparisons between recurrent selective sweeps when $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.25$ and $s = 0.01, 0.0001$. (A) The SFS of the RHH model when $s = 0.0001$ is similar to that of the standard neutral SFS whereas a large excess of singletons appears when $s = 0.01$. (B) When the SFS of the data itself is used in the *SweepFinder* calculations then the model with $s = 0.0001$ shows higher values of Λ_{MAX}. This is because the genomic regions affected by positive selection are smaller for smaller s values and a large fraction of the genome remains still unaffected by positive selection.

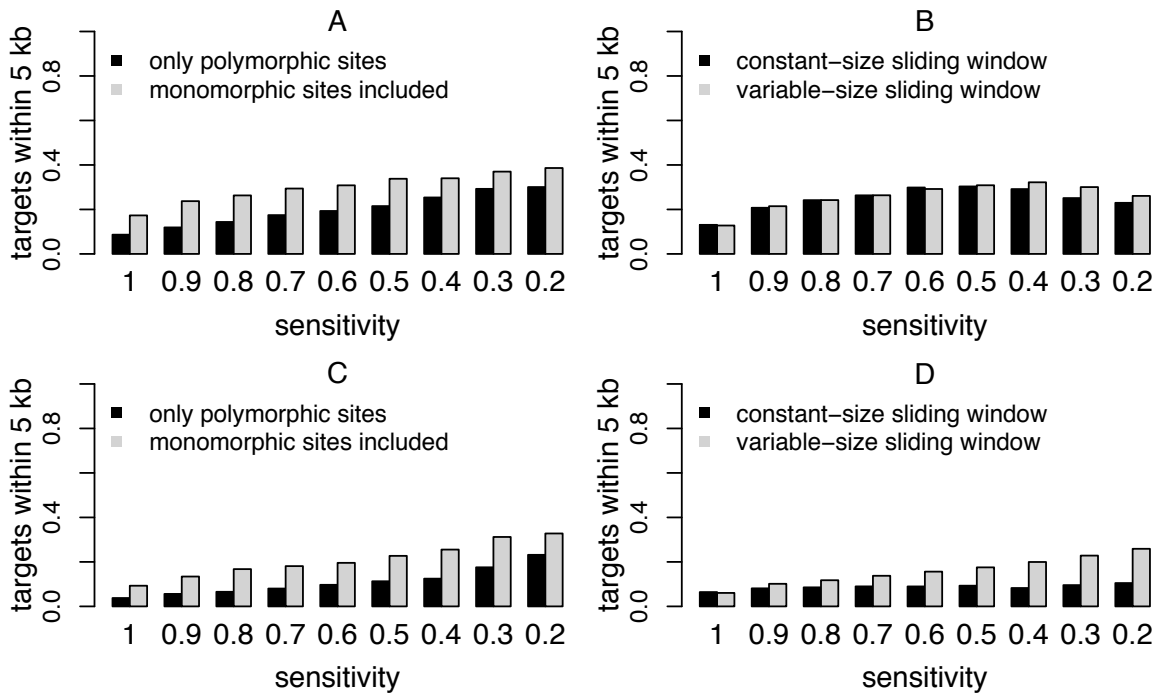

FIGURE S5.– The fraction of predicted targets within 5 kb from the true location of the selective sweep for a recurrent selective sweep scenario where $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.25$ (for A and B) and $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.50$ (for C and D). (A) and (C): Comparison of the precision of *SweepFinder* when only polymorphic sites are used (dark bars) and a fraction of monomorphic sites is embedded (light bars). (B) and (D): Comparison between the variable-size sliding window approach and the constant-size sliding window approach. The precision of the two approaches is similar for low threshold values (high sensitivity, low specificity). However, for higher cutoff values the variable-size sliding window method is slightly more precise. Simulations assume a 100-kb genomic fragment. Selective sweeps have occurred uniformly within this region or within its flanking regions following a homogeneous Poisson distribution in time. The selection coefficient is $s = 0.01$, $\theta = 0.008$/bp, and $\rho = 0.08$/bp.

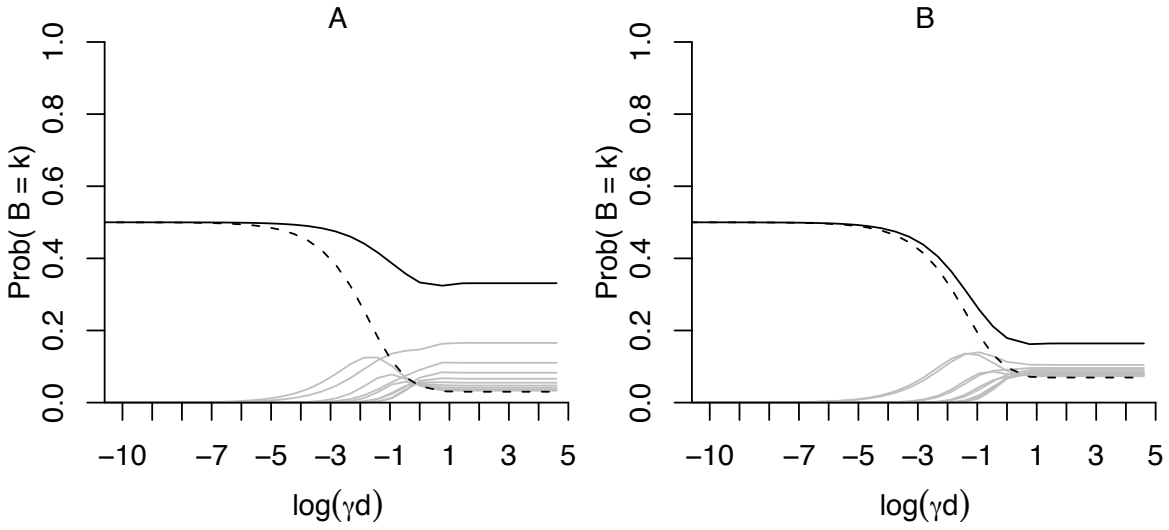

FIGURE S6.– The likelihood curves for each polymorphism class: (A) under an equilibrium selection model ($\rho = 0.05$/bp, $\theta = 0.005$/bp), and (B) under the THORNTON and ANDOLFATTO (2006) ($\rho = 0.08$/bp, $\theta = 0.008$/bp) model with selection. For both scenarios a selective sweep has occurred in the middle of a 50-kb region and the selection intensity $\alpha = 2500$. The x-axis denotes the value of parameter $\gamma = \frac{r}{s}\log(2N)$ (log scale) multiplied by the distance d from the center of the sweep. If we assume a constant recombination rate r and selection coefficient s, γ represents the distance from the location of the selective sweep x. The likelihood curve for the singletons (class '1') is depicted by the black solid line, whereas the class '11' (out of 12 sequences) is represented by a black dashed line. Gray lines illustrate the likelihood curves for the classes 2-10. For both (A) and (B) the class '1' and the class '11' contribute to the likelihood close to the sweep. Conversely, classes 2-10 contribute at larger distances from x. The major difference between (A) and (B) is that the singleton contribution is lower in (B) than (A) at larger distances. This is because the frequency of singletons is greater in (B) than in (A).

TABLE S1
The matrix used for the pre-calculation of the ω-statistic.

# SNP	$i-1$	i	$i+1$
...			
$i-1$	-	$Z_{i-1,i}$	$Z_{i-1,i}+Z_{i,i+1}$
i	-	-	$Z_{i,i+1}$
$i+1$	-	-	-
...			

TABLE S1.– A cell $Z_{i,j}$, $i < j$ represents the sum of all pairwise linkage disequilibrium comparisons (r^2) for the sites that belong to the window $[i, j]$. We have implemented a recursive algorithm in order to calculate this matrix. In detail, the calculation starts from the cell $Z_{i,i+1}$, *i.e.* the cells next to the main diagonal and proceeds upwards to the cell $Z_{i-1,i+1}$. Then $Z_{i-1,i+1} = Z_{i-1,i} + Z_{i,i+1}$. $Z_{i,i+1} = r^2_{i,i+1}$ and $Z_{i-1,i+1}$ has been calculated in the previous cycle. Then, using this matrix it is trivial to calculate the components of the ω-statistic for any configuration. When the left and right sub-regions are defined by $[i, k]$ and $[k + 1, j]$, respectively, then the numerator is the sum $Z_{i,k} + Z_{k+1,j}$ weighted by the number of calculations $[\binom{k-i+1}{2} + \binom{j-k}{2}]^{-1}$, whereas the denominator is $Z_{i,j} - Z_{i,k} - Z_{k+1,j}$ weighted by $[(k - i + 1)(j - k)]^{-1}$.