K. M. Teshima, G. Coop and M. Przeworski

SUPPLEMENTARY MATERIALS

The human model for selection on a previously neutral allele

The trajectory from frequency *f* to either fixation or its present frequency (>0) was generated as described in the Methods (under "simulation method"). However, the trajectory from *f* to loss was obtained differently, because we needed to take into account two complications: (1) Drift is more rapid in the smaller population during the bottleneck. To model this, we started at frequency *f*, and then considered the backwards diffusion process whose variance is a function of the population size. Specifically, when the frequency of the allele is *p*, the variance is given by $p(1-p)dt/x$, where *dt* is infinitesimal time interval, $x=N/N_0$, *N* is the population size at time *t* and N_0 is the reference population size, which is constant throughout the simulation. We simulated from the process unconditional on loss (the infinitesimal mean frequency change is set to 0), and retained only those trajectories that ended with absorption at 0 i.e., in which the allele was eventually lost. In other words, we only considered derived alleles. (2) More new mutations enter the large ancestral population per generation than the bottlenecked population. To take into account the change in mutational influx over time, the contribution of each backwards trajectory in which the allele is *lost* in a population of size N_0 must be weighted by a term that represents the proportion of times when the allele would *arise* in a population of size N_0. This can be done using an importance sampling method introduced by Slatkin (2001) for backward simulation of an allele in a population of changing size. He showed that to calculate the expected value of a

statistic using a backward simulation, the contribution of each sampled data set should be weighted by: $w = \frac{\Pr_F(H)}{\Pr_B(H)} N_0$,

where $Pr_F(H)$ is the probability of a sample trajectory, H, in the forward process and $Pr_B(H)$ is the probability of a sample trajectory in the backward process. In our case, $Pr_F(H) = Pr_B(H)$ so that the weight becomes $w = N_0$.

Generating the distribution of a statistic under the human model of selection on a previously neutral allele

Instead of counting the actual number of replicates that fall within a particular bin, we tabulated the sum of the importance weights for each bin, where the importance weight for the i-th replicate is N_i (N_i is the size of the population in which the allele arises in the i-th replicate). To obtain a probability distribution, we normalize the weights by dividing through by $\sum_{i=1}^{M} N_i$, where M is the total number of replicates. Once the density for a particular statistic has been generated using the importance weights, we sampled from it to generate a list of values for simulated data sets.

Supplementary materials figure legends

Figure S1: The estimated false discovery rate and false negative rate under the model for maize using π and Tajima's D. See the legend of Fig. 3 and Methods for more details. Parameters are as in Fig. 3 but the length of the simulated region is 10kb instead of 5 kb.

Figure S2: The distribution of summary statistics under a model of constant population size. The model is as described in Methods, but with b=1. The value of the summary is on the x-axis and the proportion of simulated data sets with a given value on the y-axis. The statistics presented are (a) π, (b) θ_W, (c) H, (d) Tajima's D, (e) Fu and Li's D, (f) Haplotype homozygosity (see Methods for details). The length of the simulated region is 10kb; for the other parameter values, see Methods. The light blue density is for a model of directional selection on a new mutation, the darker blue for a model of directional selection where f=0.05, the green for a model of an incomplete selective sweep in which the favored allele arose 400 generations ago and the red for the neutral model. In calculating (d) and (e), we excluded cases with no segregating sites (0.035% for of the case of selection on a new mutation, 4.5% for selection on a new, recessive mutation, 0.002% for selection on a standing variation, and 0.002% for selection starting 400 generations ago).

Figure S3: The distribution of ® and Tajima's D under four different models of selection in a human population. The value of the statistic is given on the x-axis and the proportion of simulated data sets with a given value on the y-axis. In this case, the strength of selection s = 5% and h = 0.5; other parameter values are given in the Methods. (1) The yellow density is for the case when selection acted on a previously neutral mutation (f=0.05) during the bottleneck, starting 2000 generations ago, (2) The dark blue density is for the case when selection acted on a new mutation during the bottleneck, starting 2000 generations ago, (3) The green density is for the case where selection acted on a new mutation after the bottleneck, starting 1200 generations ago,

(4) The light blue density is for the case when selection acted a previously neutral mutation (f=0.05) after the bottleneck, starting 1200 generations ago. The red histogram is for the neutral model.

Figure S4: The effect of the strength of selection on error rates. Shown are estimates of error rates using π (left two columns), Tajima's D (middle two columns) and haplotype homozygosity (right two columns) under the model for a human population. The strength of selection, s, is 1% for the 1st, 3rd and 5th columns, and 5% for 2nd, 4th and 6th columns. The false discovery rate is shown in FigS4-1 and the false negative rate in FigS4-2 (see legend of Fig.3 and Method for details). Results for the following four scenarios are presented (from top to bottom rows): (1) Selection acted on a new, co-dominant allele (h=0.5) at time τ, (2) Selection acted on a new, recessive allele (h=0.1) at time τ, (3) Selection acted on a new, co-dominant allele (h=0.5) at time 400 generations ago, (4) Selection acted on a previously neutral mutation at frequency f=0.05 at time τ.

Figure S5: The effect of matching the recombination rate in selected and neutral loci. Shown are the estimated error rates using ρ (left two columns) and Tajima's D (right two columns) under a model of selection on a new, co-dominant mutation that arose immediately after the bottleneck. The false discovery rate is shown in Figure S5-1 and the false negative rate in Figure S5-2. For the neutral loci, we considered two cases: (1) ρ is fixed to the same value as the selected loci (first and third columns), (2) ρ is chosen from the exponential distribution (second and fourth columns). ρ for the selected loci are fixed to $10\bar{\rho}$, $4\bar{\rho}$, $2\bar{\rho}$, or $\bar{\rho}$ (from top to bottom), where $\bar{\rho}$ is the median of ρ.

In this simulation, the target of selection is adjacent to the neutral region. For the other parameter values, see Methods.

Figure S6: The joint density of ® and Tajima's *D* under the model for a human population with no selection. Tajima's *D* is given on the x-axis and ® per 10kb on the y-axis. The colors denote the height of the density, with red indicating the highest probability and dark blue the lowest. The black dots indicate the values of the summary statistics for 100 simulated loci closely linked to a selected site. Selection acted on a new mutation after the bottleneck (at time τ), the dominance coefficient, *h*, is 0.5, and other parameters are as in Figure 5.

Figure S1

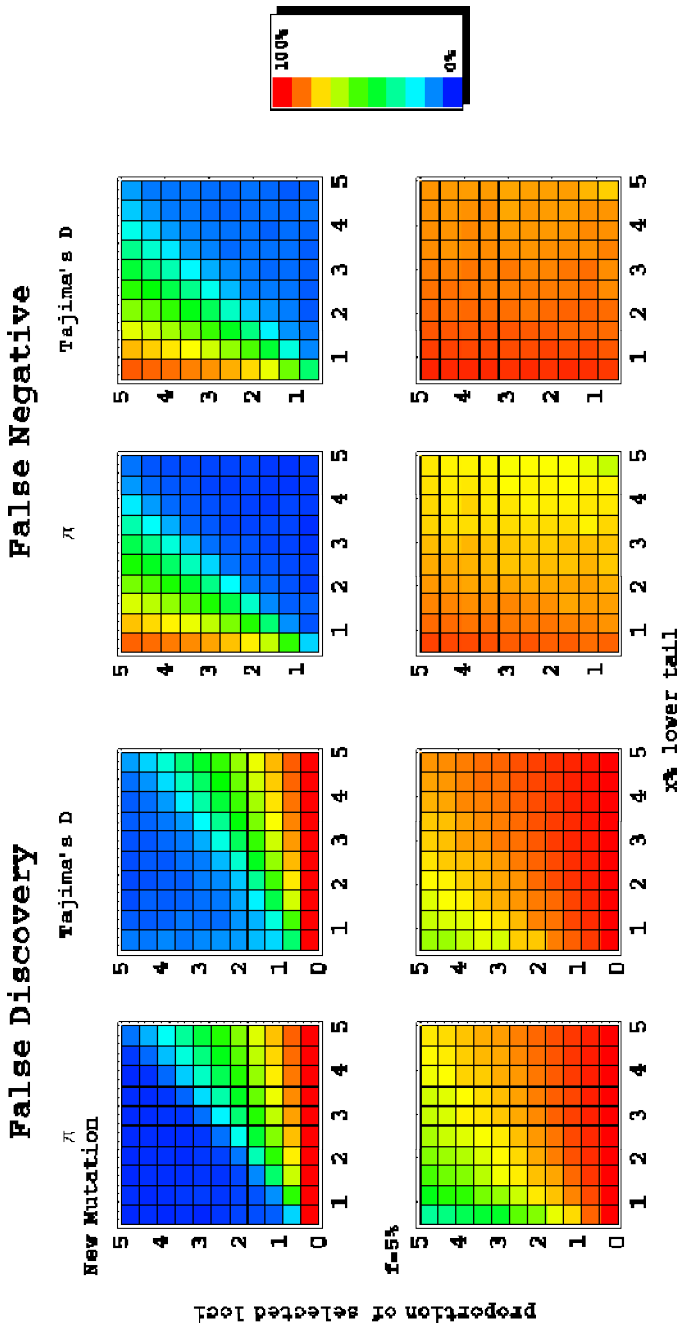

False Discovery
False Negative
New Mutation
Tajima's D
π
Tajima's D
π
100%
0%
f=5%
x% lower tail
Proportion of selected loci

Figure S2

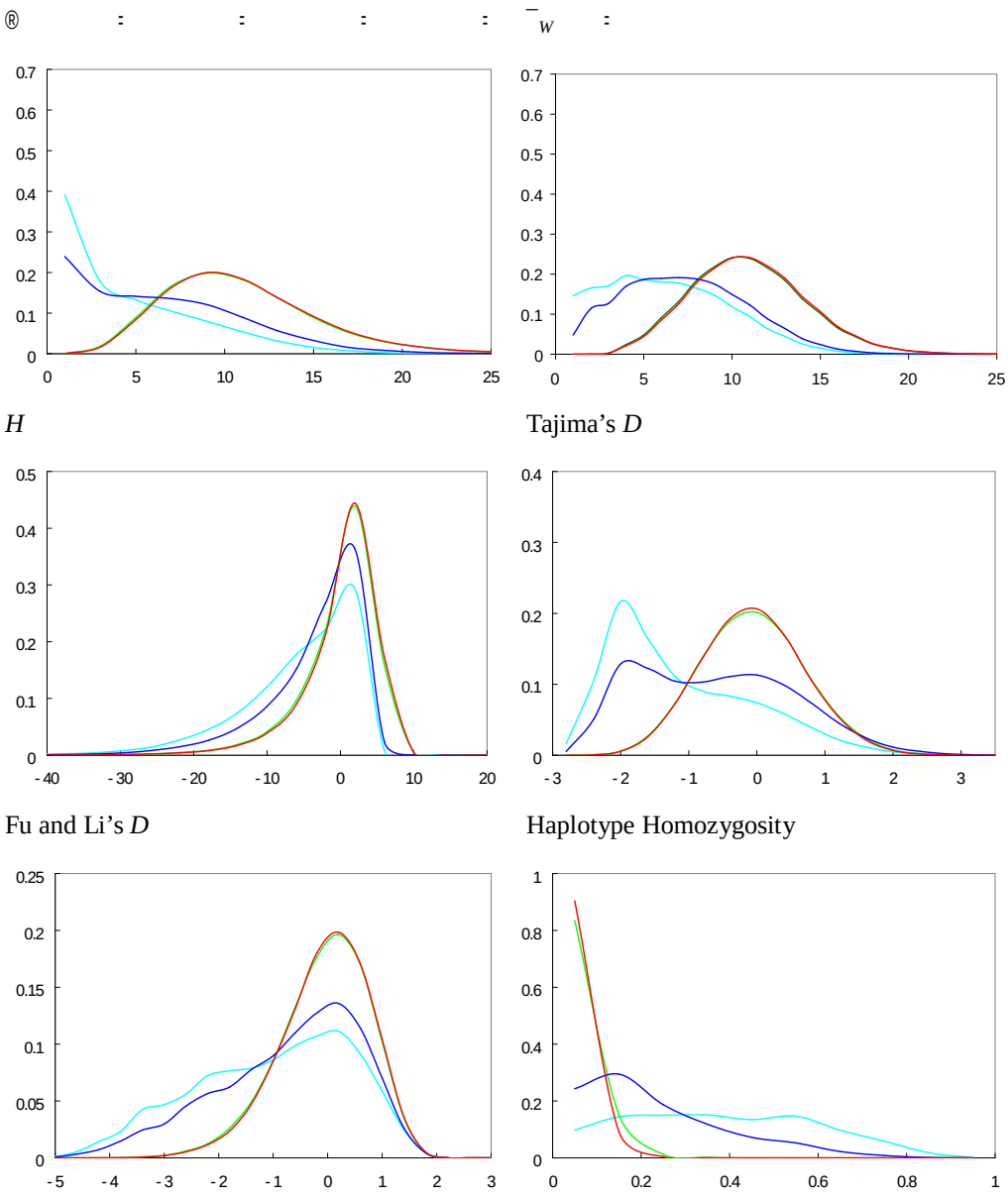
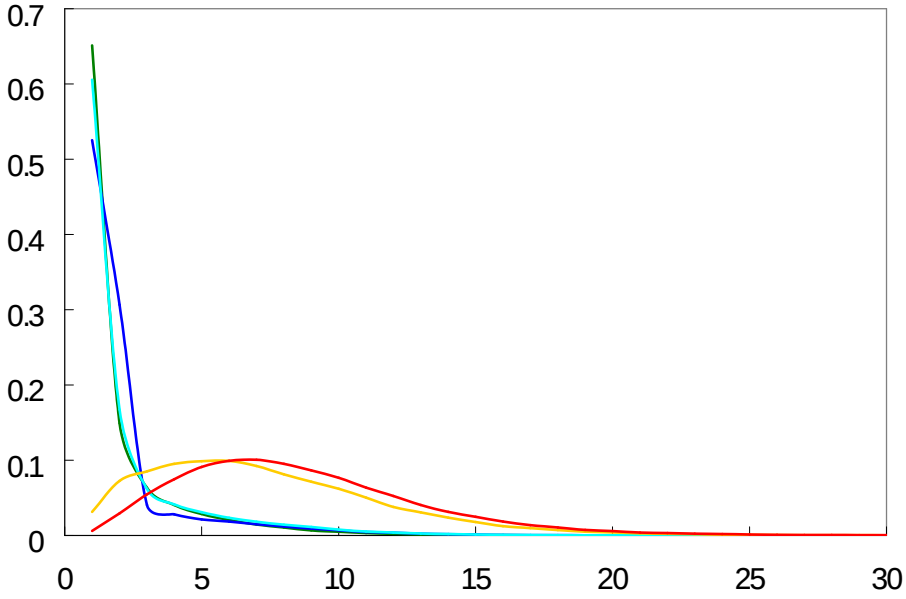

®
:
:
:
:
$\bar{w}$
:
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 5 10 15 20 25
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 5 10 15 20 25
H
Tajima's D
0.5
0.4
0.3
0.2
0.1
0
-40 -30 -20 -10 0 10 20
0.4
0.3
0.2
0.1
0
-3 -2 -1 0 1 2 3
Fu and Li's D
Haplotype Homozygosity
0.25
0.2
0.15
0.1
0.05
0
-5 -4 -3 -2 -1 0 1 2 3
1
0.8
0.6
0.4
0.2
0
0 0.2 0.4 0.6 0.8 1

Figure S3

®



Tajima's *D*

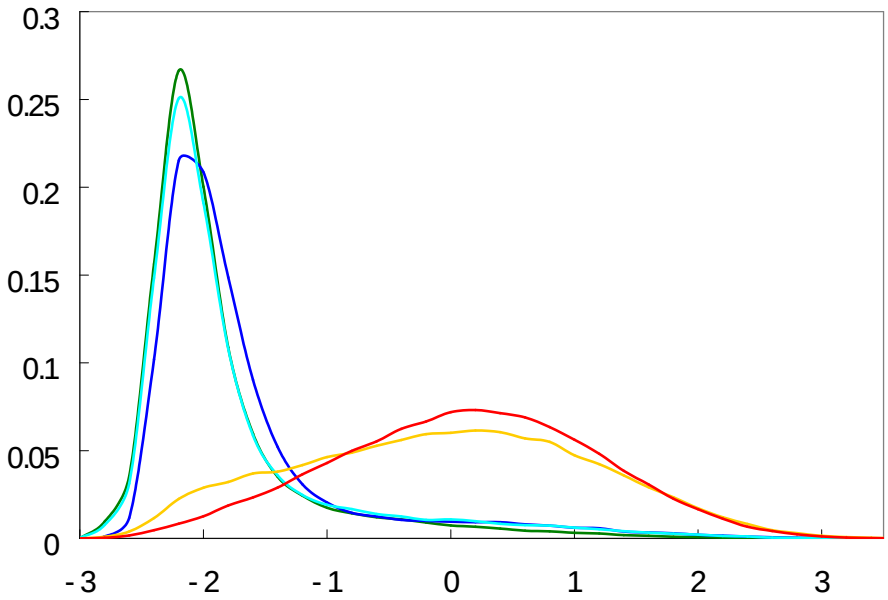


Figure S4-1

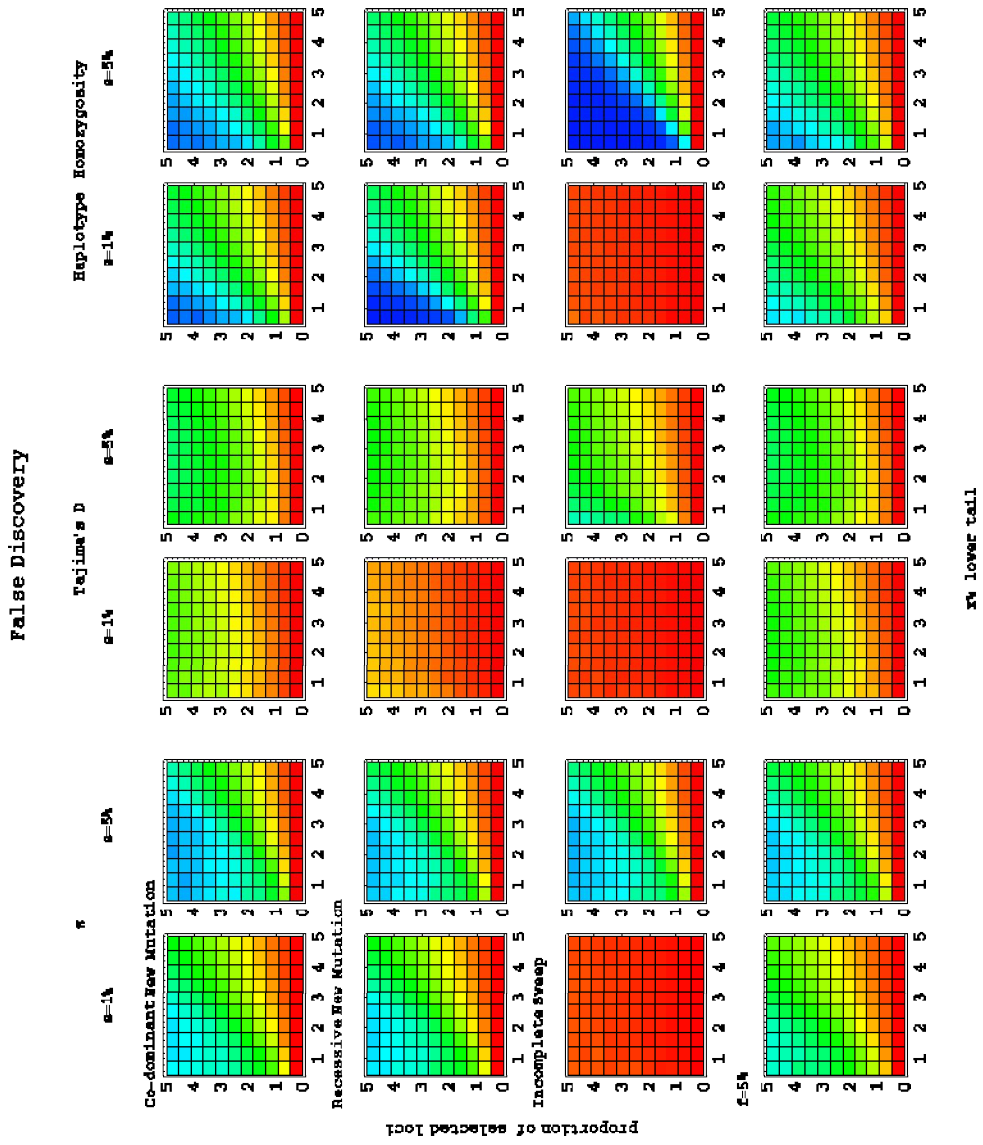

False Discovery
π
Tajima's D
Haplotype Homozygosity
s=1%
s=5%
Co-dominant New Mutation
Recessive New Mutation
Incomplete Sweep
f=5%
Proportion of selected loci
πw lower tail

Figure S4-2

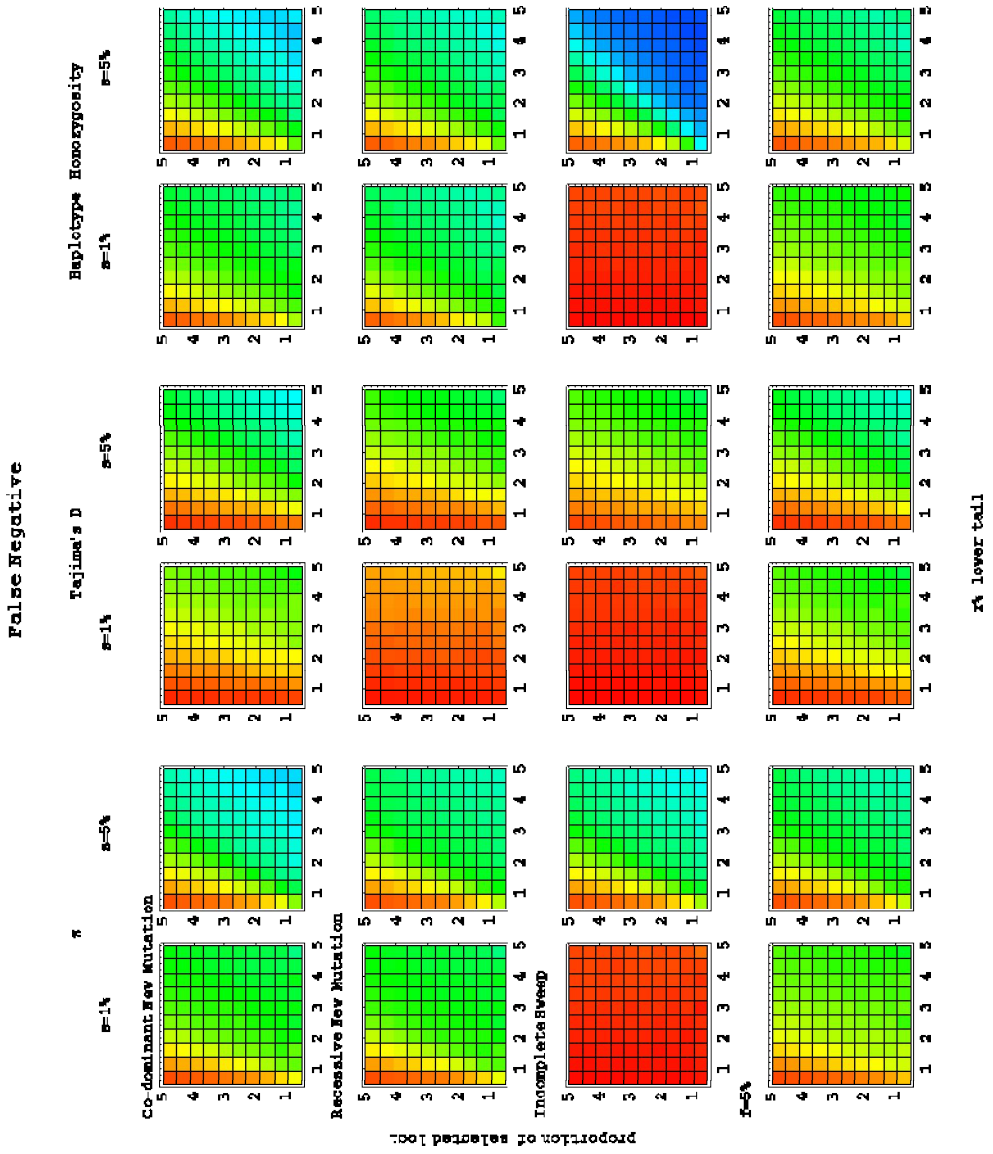

False Negative
π
Tajima's D
Haplotype Homozygosity
s=1%
s=5%
Co-dominant New Mutation
Recessive New Mutation
Incomplete Sweep
f=5%
proportion of selected loci
1% lower tail

Figure S5-1

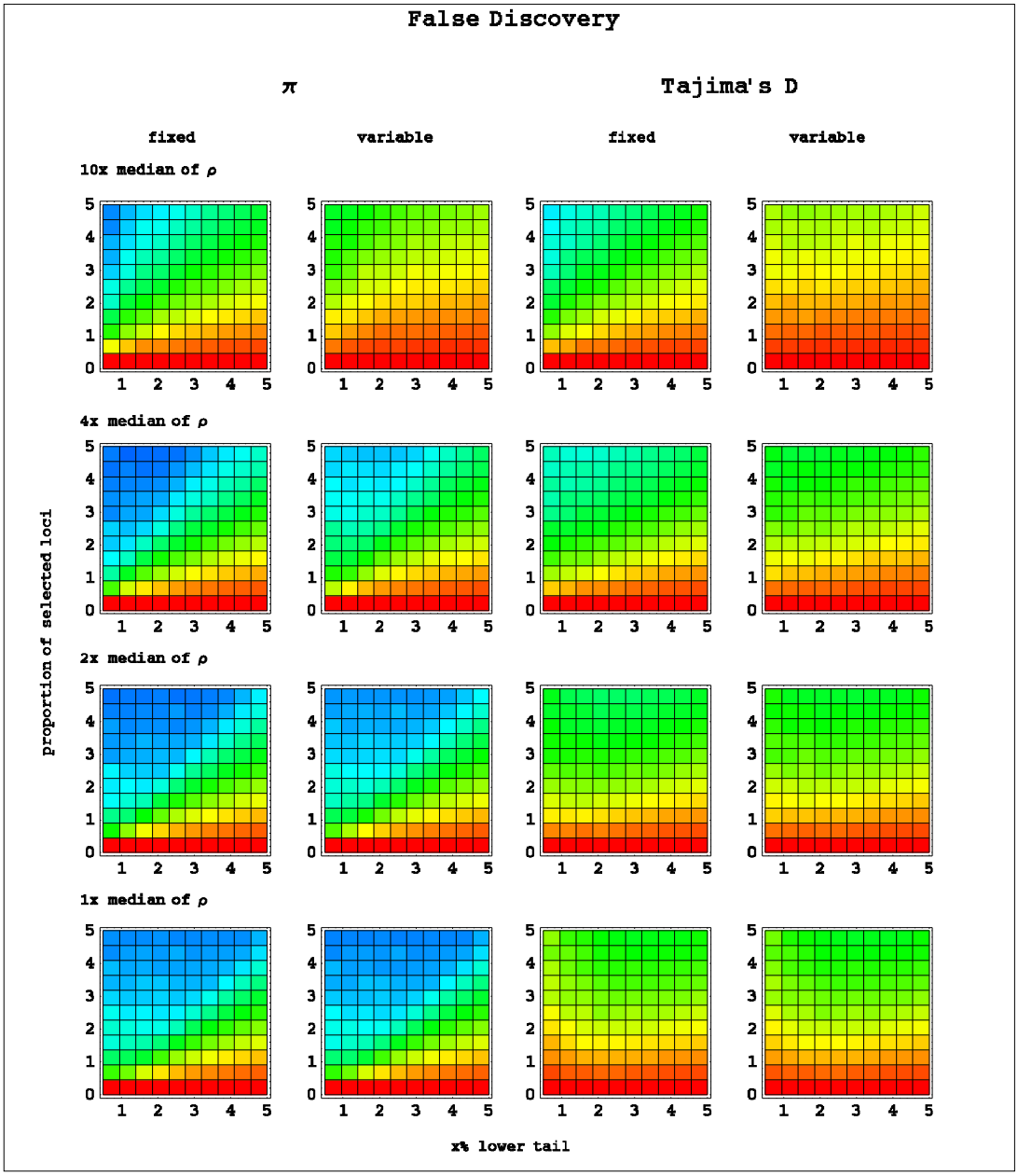

False Discovery
π
Tajima's D
fixed
variable
fixed
variable
10x median of ρ
4x median of ρ
2x median of ρ
1x median of ρ
proportion of selected loci
x% lower tail

Figure S5-2

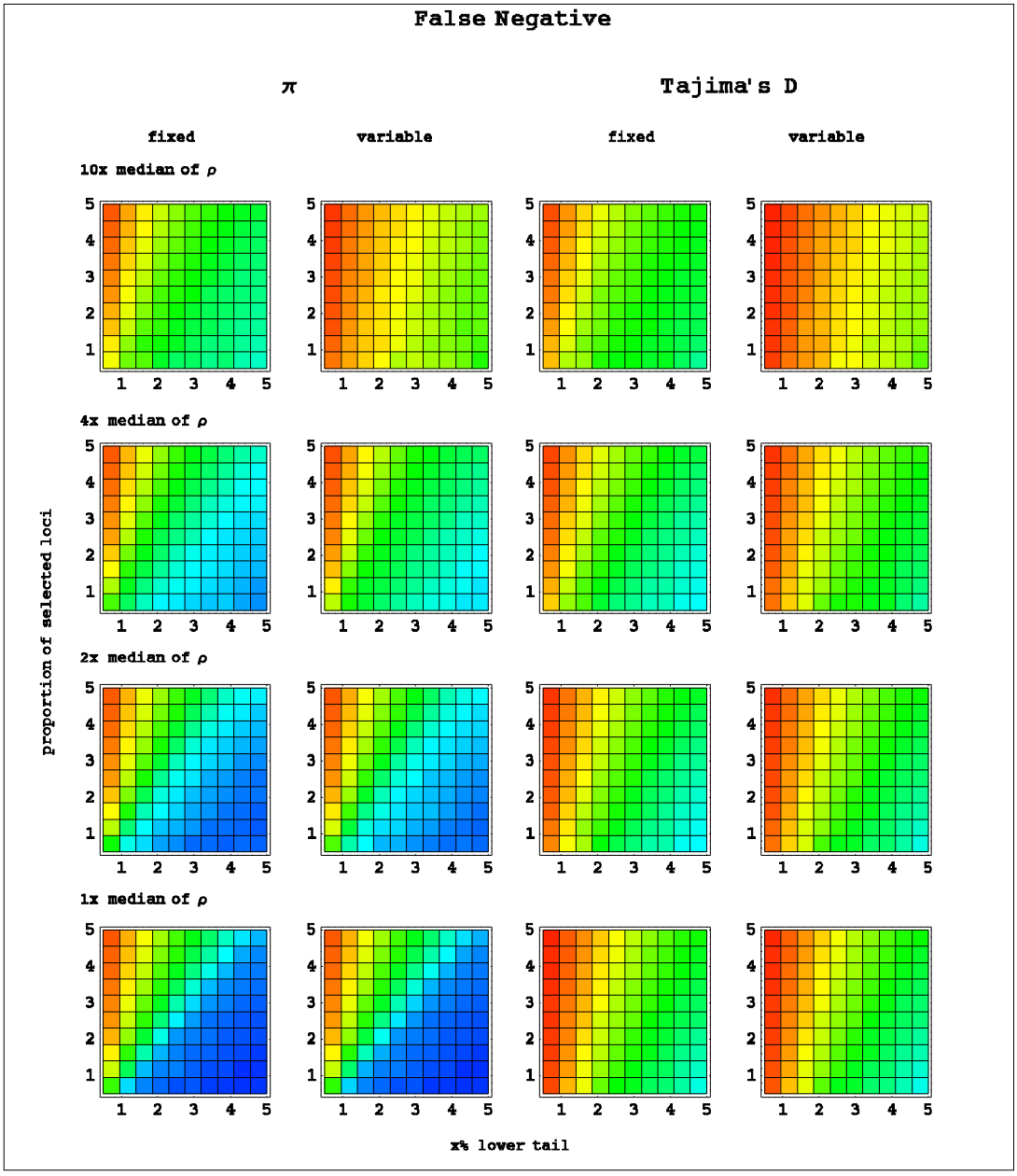

False Negative
π
Tajima's D
fixed
variable
fixed
variable
10x median of ρ
4x median of ρ
2x median of ρ
1x median of ρ
proportion of selected loci
x% lower tail

Figure S6