Simulating Tennessen et al. Model with Selection

1. Methods

Here we describe our simulation strategy for testing the ability of 'pseudo-H12' to detect hard and soft sweeps in a variety of evolutionary scenarios relevant to ancient DNA.

1.1 Demographic model

We used the Tennessen et al. demographic model which describes the ancestral human population in Africa, followed by the out of Africa event and two periods of European population growth.

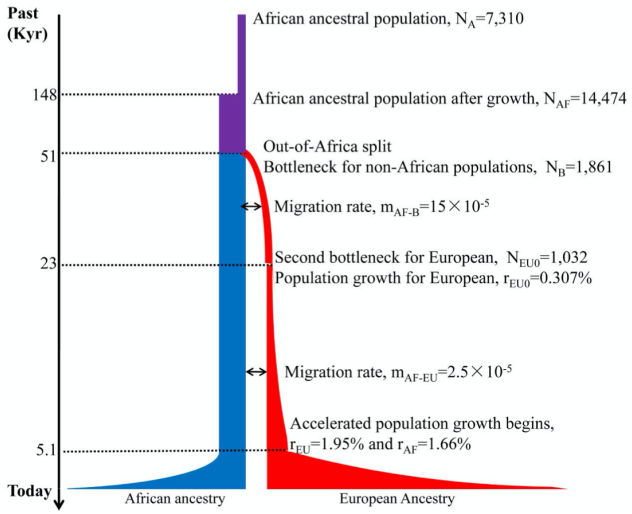

Fig1. Tennessen et al model. (Fu at al 2013, Fig. S5)

1.2 Parameters

- Mutation rate μ= 1.25e-08/bp
- Chromosome length = 5e5 bp
- Selection coefficient = ranges from 0 to 0.2
- Recombination rate = 5e-09 events/bp

For simplicity, we decided to use a constant recombination rate. We defined the recombination rate to be 5e-09 events/bp, which is a value on the lower end of the distribution of recombination rates from the DeCodeSexAveraged_GRCh36 genetic map [Kong et al, 2010]. We used a low recombination rate to avoid false positive soft selective sweeps since a high number of recombination events will likely cause hard sweeps to look "softer".

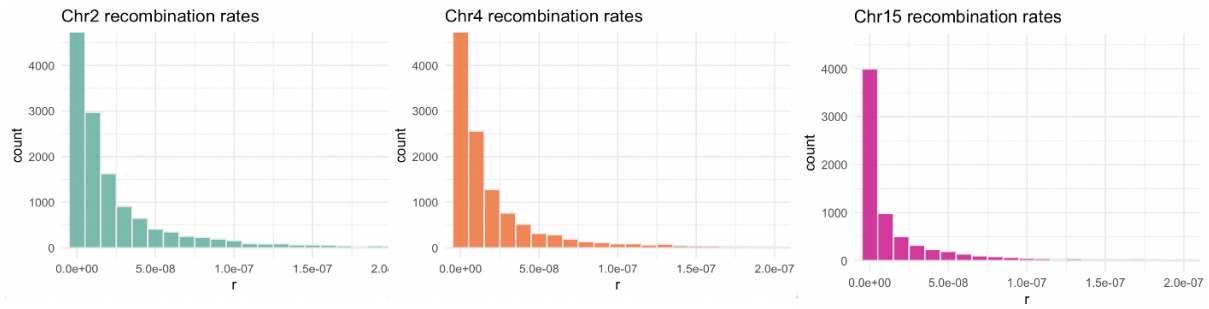

Fig 2. Recombination rates histogram for chromosomes 2,4 and 15.

1.3 Selective sweeps

We modified the *stdpopsim* code for the Two-population out-of-Africa demographic model [Tennessen et al. 2012] to include positive selection. We simulated hard sweeps varying the age of the introduced mutation and the selection coefficient. We also simulated soft sweeps by introducing K mutations *g* generations ago. We considered a mean generation time of 28 years We obtained four samples of 177 individuals each: one Mesolithic sample ~10,000 years ago (357 generations ago) and a Historical sample ~1,000 years ago (36 generations ago).
For two of our positive controls (LCT and SCL24A5), selection was likely active around 100 and 300 generations ago, for this reason, we obtain two samples from 100 and 250 generations ago.

1.3.1 Hard sweeps

We added a beneficial mutation halfway through the chromosome of a random individual from the European population. We varied the age of the mutation by introducing it 180, 280, 500, and 1000 generations ago. The introduction of the most recent mutations at 180 and 280 generations ago roughly correspond to the estimate of the onset of selection of the positive controls LCT and SLC24A5, respectively [Itan et al 2009;Wilde et al. 2014].
The simulations are conditional on the mutation not being lost, that is, we checked whether the mutation was still present in the population and restarted the simulation if it had been lost.

Running the Hard Sweep Tennessen et al. model
The SLiM code Tennessen_HardSweeps.slim requires the user to define the following parameters:

- burn_in: The burn in period is AN_e generations long. This parameter determines the constant A, commonly taken to be between 10-12.
- selec: selection coefficient of beneficial mutation.
- introduceMut: Age of the mutation or number of generations since the mutation was introduced to the European population.
- mutation_rate: mutation rate per base position per generation.
- recomb_rate: recombination rate per base position per generation.
- chromosome_length: Number of bases in chromosome.
- file_mesolithic: output file path for Mesolithic sample VCF output file.
- file_historical: output file path for Historical sample VCF output file.

An example of how to run the code is given below.

```
slim -d burn_in=10.0 -d selec=0.01 -d introduceMut=1000 -d mutation_rate=1.25e-08 -d recomb_rate=5e-09  -d chromosome_length=5e5 -d sampleSize=177 -d "file_historical='VCF_historical.csv'" -d "file_mesolithic='VCF_mesolithic.csv'"  -d "file_250='VCF_250.csv'" -d "file_100='VCF_100.csv'" Tennessen_HardSweeps.slim
```

1.3.2 Soft sweeps

We added a K=5,10,25 and 50 distinct copies of a beneficial mutation halfway through the chromosome of a random individual from the European population. We varied the age of the mutations by introducing them 1000, 500, 280 and 180 generations ago. The simulations are conditional on the mutations not being lost.

Running the Soft Sweep Tennessen et al. model
The SLiM code Tennessen_SoftSweeps.slim requires the user to define the same parameters as in the Hard Sweep model plus the following:

- K= number of copies of the mutation to introduce.
- file_mesolithic_dats: output file with data on the number of distinct copies of the beneficial mutation at when the Mesolithic sample is taken.
- File_historical_dats: output file with data on the number of distinct copies of the beneficial mutation at when the Historical sample is taken.

An example of how to run the code is given below.

```
slim -d burn_in=10.0 -d selec=0.01 -d introduceMut=1000 -d K=10  -d mutation_rate=1.25e-08 -d recomb_rate=5e-09  -d chromosome_length=5e5 -d sampleSize=177  -d "file_historical='VCF_historical.csv'" -d "file_mesolithic='VCF_mesolithic.csv'" -d "file_historical_dats='historical_dats.txt'" -d "file_mesolithic_dats='mesolithic_dats.txt'"  "  -d "file_250='VCF_250.csv'" -d "file_100='VCF_100.csv'" Tennessen_SoftSweeps.slim
```

1.4 Missing Data

Based on the missingness observed in the data, we added missing data to our simulated datasets with a mean rate of 0.54827 missingness per SNP and a standard deviation of 0.2321.

1.5 Pseudo-haplodization

We included a pseudo-haploidization scheme in which we randomly selected one of the two alleles from the heterozygous genotype as is shown below:

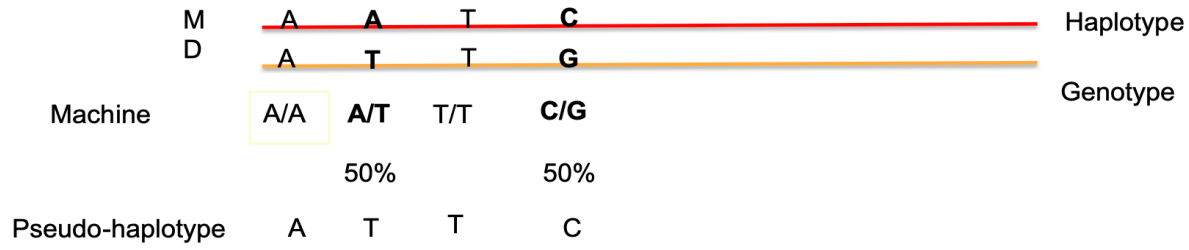

Fig3. Pseudo-haplodization scheme

1.6 Computation of pseudo-H12

The pseudo-H12 statistic is based on the haplotype homozygosity-based statistic H12 [Garud et al. 2015], which has high power to detect recent hard and soft selective sweeps. We define pseudo-H12 as:

$$\text{pseudo-H12} = (p_1 + p_2)^2 + \sum_{i>2} p_i^2,$$

where p_i is the frequency of the *i*-th most common pseudo-haplotype in a sample, obtained using the scheme in 1.5.

Missing data may inflate the pseudo-H12 statistic (**Fig. S2**) because the same haplotype is taken as a reference, which may bias the haplotypes being clustered to this specific reference haplotype. To reduce pseudo-H12 inflation resulting from missing data, we shuffle the order of the pseudo-haplotypes each time we run the code. We run the H12 code 100 times and then compute the mean of the 100 different pseudo-H12 values. However, with the high proportion of missing data modeled in the simulations, pseudo-H12 values are inflated regardless of the approach to reduce this inflation.

1.7 Sparsity of Data

To test the effect of the data sparsity on pseudo-H12 values, we simulated a chromosome 500,000 bp long, we sampled 177 individuals at four different time periods and then applied our pseudo-haplodization scheme to the sampled data. In order to take into account the sparsity of ancient DNA data, we randomly selected 201 SNPs from our pseudo-haplotype data. That is, we obtained a 201 SNP window for our sample of 177 individuals.

2. Results

We tested the ability of pseudo-H12 to detect hard and soft sweeps in a variety of evolutionary scenarios.

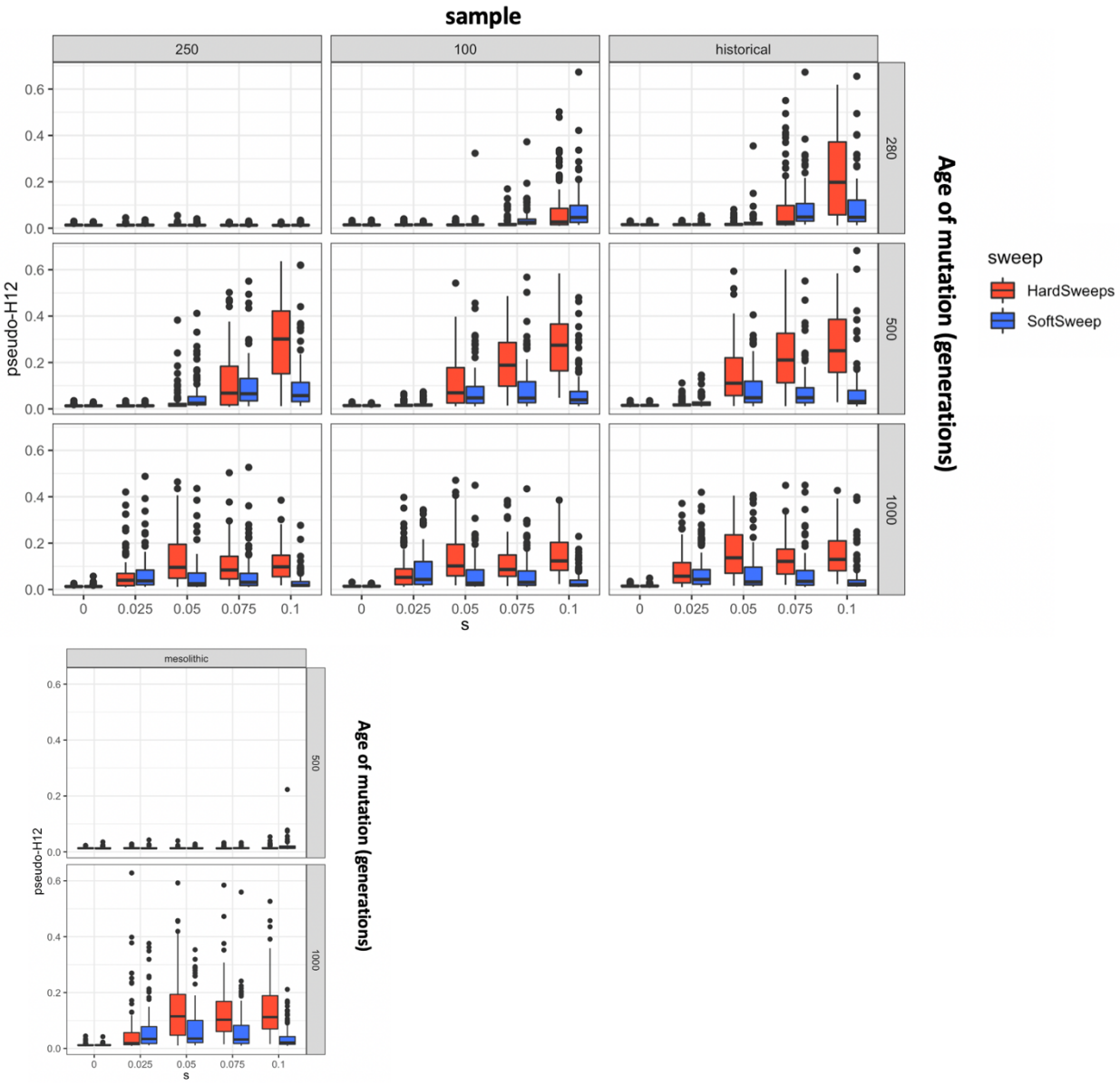

Fig 4. Pseudo-H12 values for hard sweep (red) and soft sweeps (blue) models. We show the values in the Mesolithic, 250 generations ago, 100 generations ago and Historical samples (columns) for different ages of the beneficial mutation (rows). For soft sweeps we used K=25.

Fig. 4 shows the typical pseudo H12 values computed for different samples (columns) for a range of evolutionary scenarios where we varied the selection coefficients and ages of the beneficial mutation. Without selection, the Tenneson et al. model generates low haplotype homozygosity. As selection increases, for both hard and soft sweeps, haplotype homozygosity increases for younger sweeps. As the sweep grows older (e.g. age of the mutation equals 1000 generations), recombination and mutation decay the sweep signature. In all selection cases, except when the sweep is young and selection is weak (**Fig. 4**), selection can be easily distinguished from neutrality. Hard sweeps typically give larger pseudo-H12 values.

To assess the ability of pseudo-H12 to detect sweeps of varying softness, we introduced *K* beneficial mutations *g* generations ago for *K*=5,10,25 and 50. **Fig. 5**, shows the pseudo-H12 values for different values of *K* and a selection coefficient of s=0.1. We observe that for all values of *K*, except when the sweep is young (**Fig. 5** first row, sample from 250 generations ago), selection can be easily distinguished from neutrality (*K*=0). As sweeps become softer, pseudo-H12 values become lower. When *K*=5 distinct copies of the mutation were introduced at the beginning of the sweep, the majority of the resulting sweeps were hard (**Fig S3**).

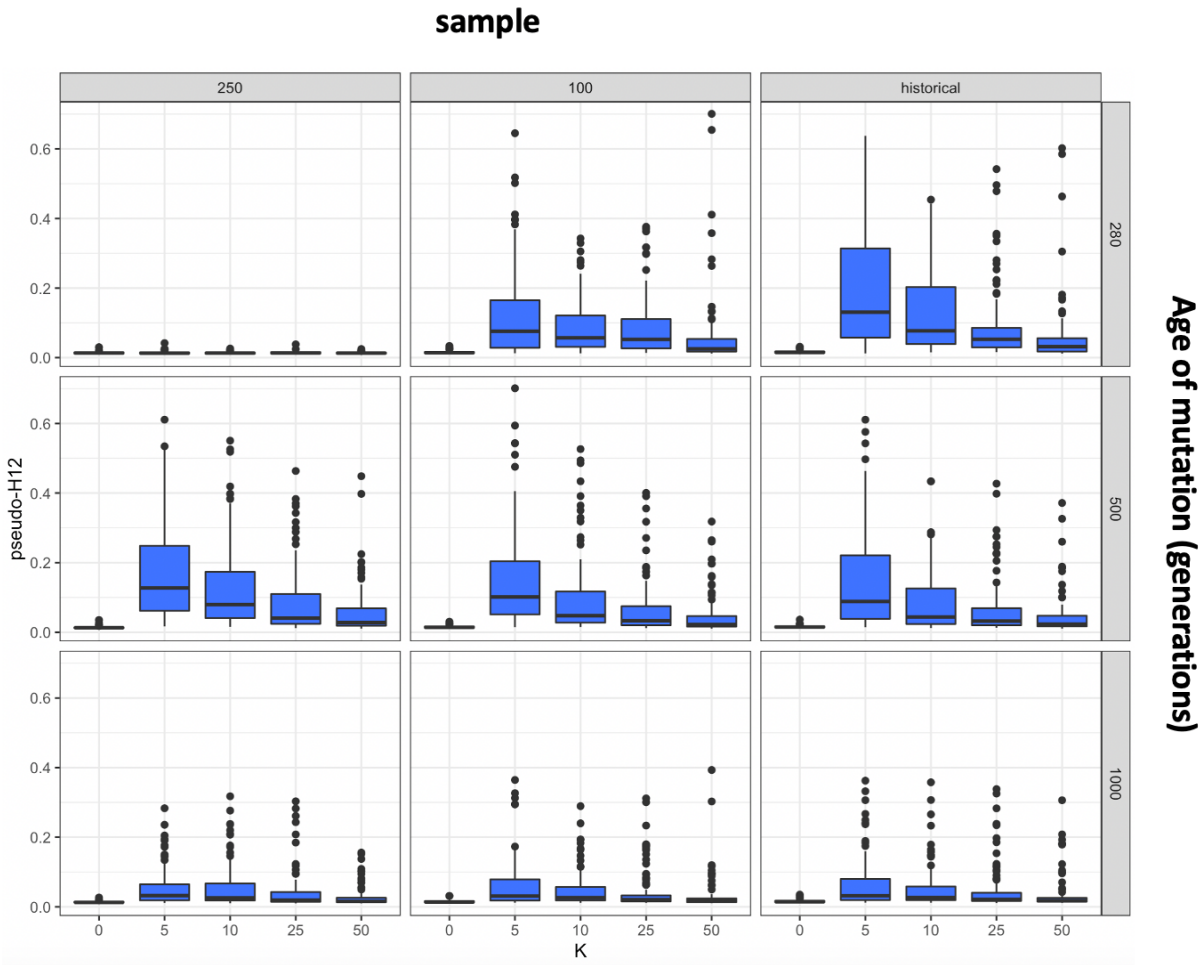

Fig 5. Pseudo-H12 values for soft sweep model. We show the values in the Mesolithic and Historical samples (columns) for different values of *K* with an average of 54.827% of missing data per site and s=0.1.

Based on these results, we conclude that pseudo-H12 has the ability to distinguish recent and strong sweeps from neutrality. Generally hard sweeps result in higher values than soft sweeps, however, we see from **Fig. 4** and **Fig. 5** that pseudo-H12 is able to distinguish both hard and soft sweeps, with exception of sweeps that are too young.

In the supplement, we test our parameter choices including:
(1) the effect of pseudo H12

(2) Missing data
(3) Softness of sweep *K*

Supplement

(1) Comparison of H12 versus pseudo-H12.

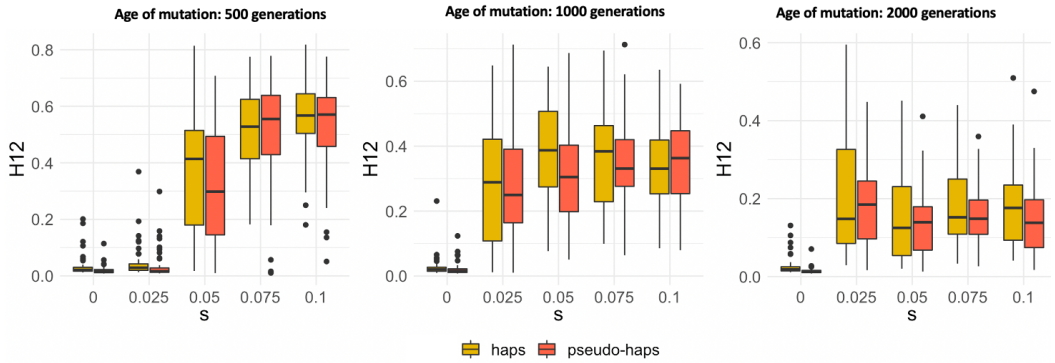

Fig S1. H12 and pseudo H12 values in Historical sample of hard sweep model with no missing data. Generally, the median value of pseudo H12 is a little lower than H12.

We also tested the impact of obtaining the average of 100 runs of H12 (**Fig S3**) and pseudo H12 (**Fig. S4**) versus using the regular method in which we run the code once. We see that, in general, the modified H12 method has a lower median H12 than the regular H12 method, making it more similar to the correct H12 value for no missing data. In this case of pseudo-H12, we don't always observe a clear decrease in the median when using the modified pseudo-H12 method and the regular pseudo-H12 method. For lower percentages of missing data, the difference between the regular and modified pseudo-H12 methods decrease (**Fig S2**).

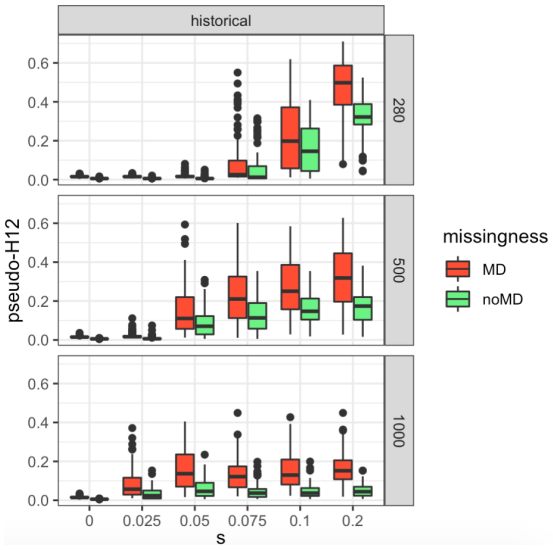

Fig S2. Pseudo-H12 values for pseudo-haplotype data from Historical sample of hard sweep model mean rate of 0.54827 missingness per SNP and a standard deviation of 0.2321

(2) Softness of Sweep

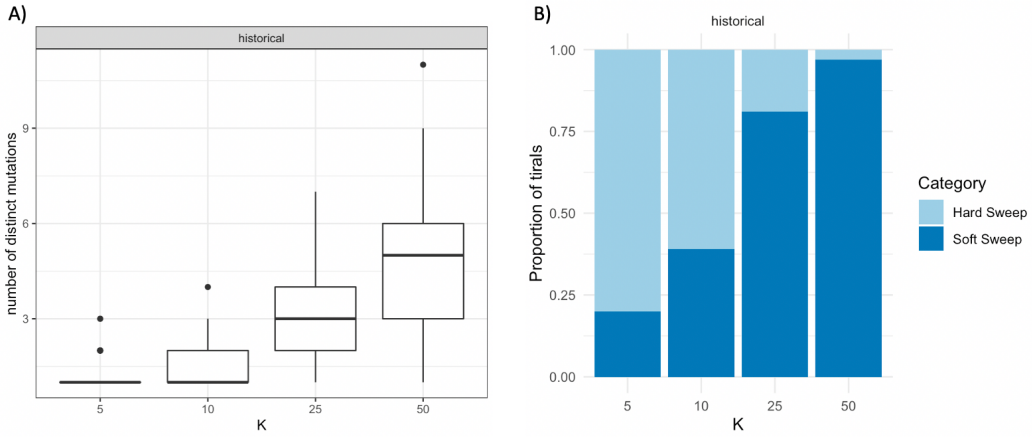

Fig S3. Softness of sweep starting with *K* distinct mutations. Figure *A)* shows the number of distinct mutations that remain in the population at the time of sampling. Figure *B)* shows the proportion of sweeps that end up being hard or soft for different starting number of mutations *K*. The figures correspond to the Historical sample of pseudo-haplotype data for *K* mutations introduced 500 generations ago and s=0.1

References

1) Adrion et al. (2020) A community-maintained standard library of population genetic models, eLife 2020;9:e54967
2) Fu, W., O'Connor, T., Jun, G. *et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493, 216–220
3) Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genetics*, *11*(2), 1–32.
4) Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The Origins of Lactase Persistence in Europe. PLOS Computational Biology 5(8): e1000491.
5) Kong, A., Thorleifsson, G., Gudbjartsson, D. *et al.* (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103.
6) Tennessen JA, et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 6;337(6090):64-9. PMID: 22604720.
7) Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, et al. (April 2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proceedings of the National A