Supplementary Information

SweeD: Likelihood-based detection of selective sweeps in thousands of genomes

Pavlos Pavlidis[1*], Daniel Živković[2], Alexandros Stamatakis[1], Nikolaos Alachiotis[1]
[1]The Exelixis Lab, Scientific Computing Group, Heidelberg Insitute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, D-69118, Heidelberg, Germany
[2]Section of Evolutionary Biology, Biocenter, University of Munich, D-82152 Planegg-Martinsried, Germany

S1. Command-lines for the simulations that were used to compare the performance of SweeD and SweepFinder

We have used the software msms (Ewing and Hermisson 2010) to simulate neutral datasets and datasets with selection. The command line for the neutral dataset is:

```
java -Xmx125000M -jar msms.jar -N 1000000 -ms X 1 -s S -r 5000 5000
```

where **X** is the sample size and **S** denotes the number of SNPs. In our case **X** = {50, 100, 750, 1000} and **S** = {10000, 100000, 1000000}.

For the dataset with selection the command line is:

```
java -Xmx125000M -jar msms.jar -N 1000000 -ms X 1 -s S -r 5000 5000 -Sp 0.5 -SF 0.001 1.0 -SAA 10000
-SAa 5000 -Saa 0
```

S2. Command lines for the simulations that were used to assess the effect of sample size on the accuracy of sweep detection

We have used the following command lines for simulating a sample from a constant population:

```
 msms -N 1000000 -ms X 1000 -t 4000 -r 10000 40000 -Sp 0.5 -STrace all_trajectory.txt -oSweeD -length 400000
```

-N 1000000 denotes the present-day population size
-ms X 1000 denotes that X sequences were simulated 1000 times. X = 12, 50, 100, 500, 1000
-t 4000: the value of the θ parameter for the whole region
-r 10000 40000: the value of ρ:=10000 and there are 40000 points on the genome where recombination may occur
-Sp 0.5 denotes that the recombination has occurred at the middle of the genomic regions
-STrace all_trajectory.txt specifies the file where the trajectories of the beneficial allele have been stored
-oSweeD: output data in SF format
-length 400000: the length of the genomic region

For sampling from a bottlenecked population that was described in the main text, we have used the following command-line:

```
msms -N 100000 -ms X 1000 -t 4000 -r 10000 40000 -Sp 0.5 -STrace all_trajectory.txt -oSweeD -length 400000 -eN 0.0375 0.01 -eN 0.03875 1.0
```

The flags are the same as in the constant-size model. The flags -eN 0.0375 0.01 -eN 0.03875 1.0 specify the demographic model:
-eN 0.0375 0.01: at time 0.0375 (scaled in units of 4N) pastwards, the population was decreased by a factor of 100.
-eN 0.03875 1.0: at time 0.03875 (scaled in units of 4N) pastwards, the population size was as large as in the present-day.

The three additional bottleneck models that we used to further generalize the results for the bottleneck model can be simulated with the following command lines:

Excess of intermediate- and high-frequency derived alleles

```
msms -N 1000000 -ms X 1000 -t 4000 -r 10000 40000 -Sp 0.5 -STrace all_trajectory.txt -oSweeD -length 400000 -eN 0.025 0.01 -eN 0.027 1
```

Excess of low- and high-frequency derived alleles

```
msms -N 1000000 -ms X 1000 -t 4000 -r 10000 40000 -Sp 0.5 -STrace all_trajectory.txt -oSweeD -length 400000 -eN 0.025 0.01 -eN 0.035 1
```

Excess of low-frequency derived alleles

```
msms -N 1000000 -ms X 1000 -t 4000 -r 10000 40000 -Sp 0.5 -STrace all_trajectory.txt -oSweeD -length 400000 -eN 0.025 0.01 -eN 0.075 1
```

Note that X = 12, 50, 100, 500 for the last three models in order to save computational time.

S3. Additional bottleneck models to assess the effect of the sample size on the accuracy of sweep localization.

We have simulated three additional bottlenecks to further generalize the results regarding the effect of sample size on the accuracy of sweep localization. The parameters of these bottleneck models have been chosen to generate site frequency spectra that are characterized by i) excess of low-frequency derived alleles, ii) excess of low- and high-frequency derived alleles, and iii) excess of intermediate- and high-frequency derived alleles. Figure S1 shows the site frequency spectra for a sample of 20 sequences. The Hudson's ms command lines flags that correspond to the demographic models are:

1. -eN 0.025 0.01 -eN 0.027 1
2. -eN 0.025 0.01 -eN 0.035 1
3. -eN 0.025 0.01 -eN 0.075 1

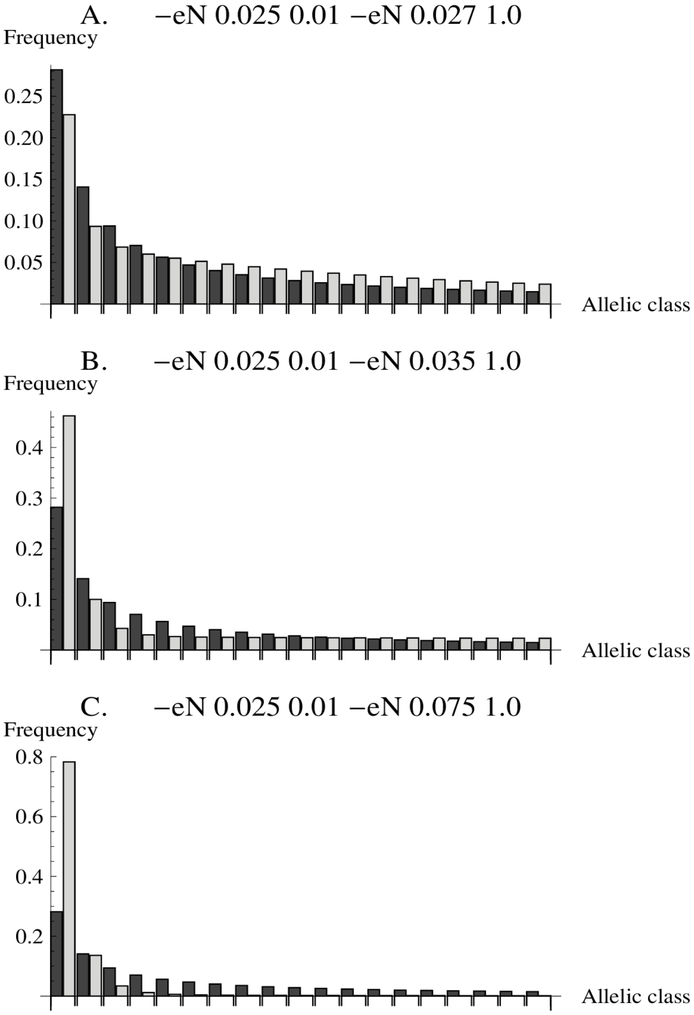

Figure S1: The site frequency spectra of three bottleneck models (light bars) that were used to assess the effect of sample size on the accuracy of sweep detection. We also show the SFS of the standard neutral model (dark bars) for comparison.

For all three bottleneck models increasing sample size results in greater accuracy of sweep localization (Figure S2).

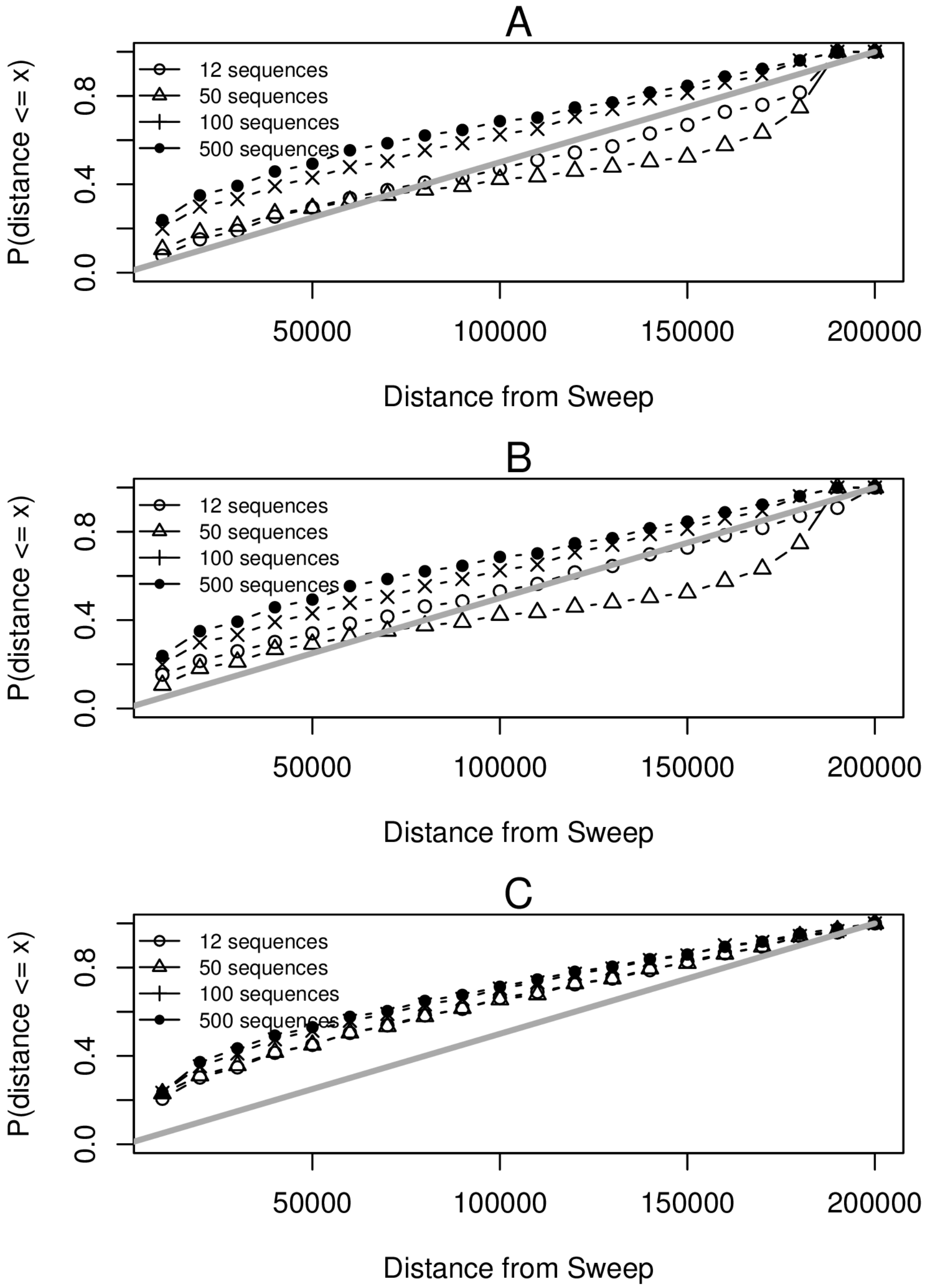

Figure S2: Assessment of the accuracy of predicting the selective sweep position for various sample sizes and demographic parameters. The x-axis in all plots shows the distance *d* of the reported selective sweep position from the true selective sweep position. Distance is grouped in bins of size 10000, i.e. d_1 = 10000, d_2 = 20000, …, d_{20} = 200000. For each bin *i,* the y-axis shows the frequency of simulated datasets with a reported selective sweep position at a distance less than d_i. The plots refer to the same bottleneck models as in Figure S2. In detail, the demographic model of (A) is described by Hudson's ms command line options -eN 0.025 0.01 -eN 0.027 1. In (B) the respective Hudson's ms command line options are -eN 0.025 0.01 -eN 0.035 1, and in (C) -eN 0.025 0.01 -eN 0.075 1. The straight line depicts the expected percentage of simulations at each bin, if the position of a reported selective sweep would be distributed uniformly along the simulated fragment of 400 kb. The figure shows that accuracy of detecting selective sweeps increases with sample size in all bottlenecked populations.

S4. Genes located in the outlier genomic regions

Supplementary Table S1

Description of the outlier genes (significance threshold at 0.01 for *both* SweeD and OmegaPlus) of the chromosome 1 of the 1000 Genomes dataset. Columns OmegaPlus and SweeD provide the score for the OmegaPlus and SweeD, respectively. The column Function provides a summary of the function of the gene. Gene Name and Gene Description give the gene name and a short description of the gene.

Gene Name	**OmegaPlus**	**SweeD**	**Gene Description**	**Function**
IFFO2	2086.949219	14.72967	intermediate filament family orphan 2	-
PIGV	1465.577637	33.22361	phosphatidylinositol glycan anchor biosynthesis, class V	Alpha-1,6-mannosyltransferase involved in glycosylphosphatidylinositol-anchor biosynthesis. Transfers the second mannose to the glycosylphosphatidylinositol during GPI precursor assembly
FAM46B	437.057983	8.575692	family with sequence similarity 46, member B	-
WDTC1	312.270355	19.43544	WD and tetratricopeptide repeats 1	enzyme inhibitor activity, protein binding, histone binding
CSMD2	1528.953003	21.42344	CUB and Sushi multiple domains 2	
BEND5	270.87326	11.6375	BEN domain containing 5	
AGBL4	1651.128296	9.678677	ATP/GTP binding protein-like 4	Metallocarboxypeptidase that mediates deglutamylation of target proteins. Catalyzes the deglutamylation of polyglutamate side chains generated by post-translational polyglutamylation in proteins such as tubulins. Also removes gene-encoded polyglutamates from the carboxy-terminus of target proteins such as MYLK. Acts as a long-chain deglutamylase and specifically shortens long polyglutamate chains, while it is not able to remove the branching point glutamate, a process catalyzed by AGBL5/CCP5 (By similarity)
ELAVL4	1886.829468	10.29922		May play a role in neuron-specific RNA processing. Protects CDKN1A mRNA from decay by binding to its 3'-UTR (By similarity). Binds to AU-rich sequences (AREs) of target mRNAs, including VEGF and FOS mRNA
NFIA	1390.370483	17.82784	nuclear factor I/A	Recognizes and binds the palindromic

				sequence 5'-TTGGCNNNNNGCCAA-3' present in viral and cellular promoters and in the origin of replication of adenovirus type 2. These proteins are individually capable of activating transcription and replication
DEPDC1	355.366943	18.8921	DEP domain containing 1	May be involved in transcriptional regulation as a transcriptional corepressor. The DEPDC1A-ZNF224 complex may play a critical role in bladder carcinogenesis by repressing the transcription of the A20 gene, leading to transport of NF-KB protein into the nucleus, resulting in suppression of apoptosis of bladder cancer cells
SRSF11	1691.930786	8.277768	serine/arginine-rich splicing factor 11	This gene encodes 54-kD nuclear protein that contains an arginine/serine-rich region similar to segments found in pre-mRNA splicing factors. Although the function of this protein is not yet known, structure and immunolocalization data suggest that it may play a role in pre-mRNA processing. Alternative splicing results in multiple transcript variants encoding different proteins. In addition, a pseudogene of this gene has been found on chromosome 12.(provided by RefSeq, Sep 2010) May function in pre-mRNA splicing
ZNF326	2922.085205	10.47725	zinc finger protein 326	Core component of the DBIRD complex, a multiprotein complex that acts at the interface between core mRNP particles and RNA polymerase II (RNAPII) and integrates transcript elongation with the regulation of alternative splicing: the DBIRD complex affects local transcript elongation rates and alternative splicing of a large set of exons embedded in (A + T)-rich DNA regions. May play a role in neuronal differentiation and is able to bind DNA and activate expression in vitro
RPL5	916.188721	10.76833	ribosomal protein L5	Required for rRNA maturation and formation of the 60S ribosomal subunits. This protein binds 5S RNA
MTF2	566.222595	8.334026	metal response element binding transcription factor 2	Binds to the metal-regulating-element (MRE) of metallothionein-1A gene promoter. Binding is zinc-dependent (By similarity)
DPYD	441.677795	15.29442	dihydropyrimidine	Involved in pyrimidine base degradation.

			dehydrogenase	Catalyzes the reduction of uracil and thymine. Also involved the degradation of the chemotherapeutic drug 5-fluorouracil
FLJ39739	302.900482	9.106405	Uncharacterized **FLJ39739**	
S100A11	8305.124023	15.14509	S100 calcium binding protein A11	Facilitates the differentiation and the cornification of keratinocytes
GON4L	499.727264	22.39637		
LOC440704	2182.756104	17.56356		
KCNT2	690.009888	10.04751	potassium channel, subfamily T, member 2	Outward rectifying potassium channel. Produces rapidly activating outward rectifier K(+) currents. Activated by high intracellular sodium and chloride levels. Channel activity is inhibited by ATP and by inhalation anesthetics, such as isoflurane (By similarity). Inhibited upon stimulation of G-protein coupled receptors, such as CHRM1 and GRIA1

S5. Deviation of the simulation-based estimation of the SFS from the theoretical expectation

A flexible approach to estimate the average SFS for a variety of neutral demographic scenarios is by using simulations. For example Hudson's ms and msms (Ewing and Hermisson 2010) allow for simulating a multitude of demographic scenarios (with one or more demes, constant or varying population size, etc), and thus it enables the estimation of the neutral average SFS for a variety of demographies. A disadvantage of using simulations is that a large number of replications is required to acquire as accurate results as the analytical calculations (using the MPFR library for arbitrary precision).

We estimated the accuracy of the SFS calculations as follows. We assumed a sample of size n=1000 sequences from a bottlenecked population. The population size was decreased by a factor of 100 at time 0.0375 to attain its present-day level at time 0.03875 (time is measured in units of 4N generations and proceeds backwards; 0.0375 corresponds to 150,000 generations). The theoretical expected relative frequency $r_{1000,i}$ (here abbreviated as r_i) of a SNP of fr*equ*ency i as given in the main text was calculated using Mathematica (v. 7). Let S_i be the frequency of a SNP of fr*e*quency i estimated via simulations. For a SNP of fr*e*quency i, de*vi*ation D_i was defined as:

$$D_i = \begin{matrix} S_i / r_i - 1, \mathit{for}\, S_i > r_i \\ r_i / S_i - 1, \mathit{for}\, S_i < r_i \\ 0, \mathit{otherwise} \end{matrix} .$$

Thus, D_i is positive for SNPs whose frequencies are overestimated and negative otherwise. Results are shown in Figure S3. Note that deviations are of the order of 0.1 when the average SFS is calculated using 1000 simulated datasets.

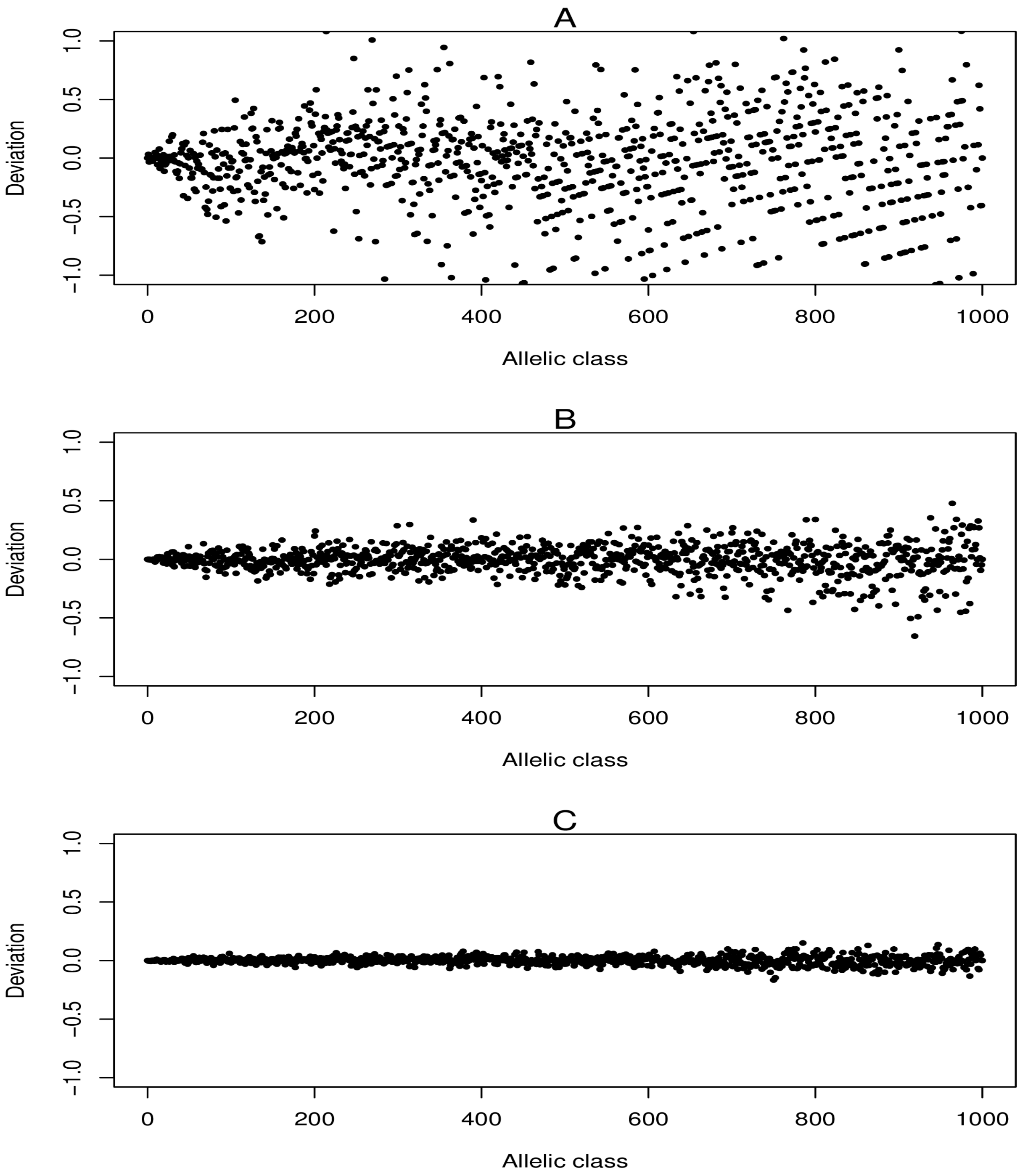

Figure S3: Deviation of thesimulated average SFS from the theoretical expectation. A bottleneck scenario was used for the simulations. A) 10 datasets were simulated to calculate the average SFS. Absolute deviation values are larger than 1, but for illustration purposes we show values between -1 and 1 only. B) 100 datasets were simulated to calculate the average SFS and C) 1000 datasets were simulated. Command-lines are provided in the supplementary section **"Command-lines for simulations used in comparing the time needed to estimate the SFS either by simulations or by the SweeD".** Deviation decreases with the number of simulated datasets. Furthermore, deviation is larger for high-frequency derived SNPs.

S6. Command-lines for simulations used in comparing the time needed to estimate the SFS either by simulations or by the SweeD

The SFS was estimated either by SweeD (using the MPFR library) or by simulations (using msms). The following command line was used for SweeD:

```
SweeD -name $X -s $X -osfs SweeD_$X.SFS -eN 0.0375 0.01 -eN 0.03875 1.0,
```

where $X = 10, 100, 500, 1000 denotes the sample size.

For simulations we used the commands:

```
msms -ms $X $R -t 1000 -r 1000 2000 -eN 0.0375 0.01 -eN 0.03875 1.0
```

where $X = 10, 100, 500, 1000 denotes the sample size,
$R = 10, 100, 1000 denotes the number of replications.

For the recombination rate, we used the value 1000, which resembles approximately the recombination rate of a fragment of 10 kb in *D. melanogaster* (European population). We could use a lower recombination rate (which would accelerate the simulations). However, using a lower value for the recombination rate would require a higher number of simulations in order to produce results of similar accuracy.