Founder event characterization for ancient samples of a diploid species, considering only the 5 five chromosomes

In this example, we are using data simulated under a simple population model where a target population (*Pop1*) with effective size N_0=12,500 experienced a founder event 50 generations before present (gBP) with intensity 10%. This population diverged from another one (*OUTGROUP*) 1,800 generations ago. We will consider that the samples are from ancient DNA and thus have to be pseudo-haploidized (i.e. only one read at the SNP position was sampled for the individual and set as homozygous).

Data

The data must be in EIGENSTRAT format (see here for a description of the format) and consists of 3 files:

- `data/example.geno.gz` is a gzipped geno file,
- `data/example.snp.gz` is a gzipped snp file,
- `data/example.ind` is the ind file.

Decomposition of the *ASCEND* parameter file

We now generate a parameter file:

Location of the geno file
Note that the file can be gzipped or not.

```
genotypename: example.geno.gz
```

Location of the snp file
Note that the file can be gzipped or not.

```
snpname: example.snp.gz
```

Location of the ind file

```
indivname: example.ind
```

Prefix of the output files
Can also be an absolute or relative path which includes directories. If the directory(ies) do(es) not exist, *ASCEND* will create it.

```
outputprefix: results/test
```

Indicate specific chromosomes to analyze
Sometimes, we may want to limit the analysis to specific chromosomes. Here, for instance, we consider only the first 5 chromosomes:

```
chrom: 1, 2, 3, 4, 5
```

If you do not want to subset your data to specific chromosomes, do not provide this option.

Name of the target population on which to carry the estimation

```
targetpop: Pop1
```

Name of the outgroup population
In this example where we work on ancient DNA, we advice not subtracting the cross-population allele sharing using an outgroup population, since it is challenging to find a proper outgroup with same sample age as the target population. Thus we set the option as *None* (alternatively, we can also comment the option by adding a # in the beginning).

```
outpop: None
```

Binning size (in Morgans) of the genetic distances
We advice using 0.001 Morgans.

```
binsize: 0.001
```

Minimum genetic distance (in Morgans) to consider
We advice starting at 0.001 Morgans.

```
mindis: 0.001
```

Maximum genetic distance (in Morgans) to consider
We advice calculating the allele sharing correlation up to 30 cM but consider increasing this value if you have intuition of a very recent founder event or if visual inspection of the decay curve calculated on a first run has not decayed to 0 for the higher distances.

```
maxdis: 0.3
```

Maximum proportion of missing allele sharing values
We suggest setting this as 1.0. We will exclude all SNPs with a proportion of missing values greater or equal to this proportion.

```
maxpropsharingmissing: 1
```

Minimum minor allele frequency
Minimum MAF of a SNP for it to be considered in the analysis. We suggest setting it as 0. Any SNP with a MAF lower or equal to this value will be excluded.

```
minmaf: 0
```

Ploidy
If your individuals are haploid, set *NO*, otherwise set *YES*. Here the genotypes are diploid, so we set:

```
haploid: NO
```

Specify if the genotypes are pseudohaploid or not

For ancient DNA, genotypes are often pseudohaploid (or if not already provided as pseudohaploid in the input *.snp file, they have to be pseudohaploidized). Setting the option as *YES* will also automatically switch the *ASCEND* algorithm to use allele sharing weighted covariance (instead of correlation) to provide a less biased estimation of founder intensity.

```
dopseudohaploid: NO
```

Specify the unit of the genetic distances

If the genetic positions provided in the *.snp file are in Morgans, set *YES*, else set *NO*. In the dataset we are using here, distances are given in cM (you can check looking at the last line of the example.snp.gz file that the genetic distance is around 100 times the physical position times the human mean recombination rate of 10^{-8}).

```
morgans: NO
```

Say if you only want to do the fitting on already generated *.perchrom.outs file

```
onlyfit: NO
```

Algorithm for the calculation of allele sharing correlation

We **strongly** recommand using fft. If `usefft: NO`, we will use the (slow) Naive algorithm.

```
usefft: YES
```

Number of sub-bins for FFT calculation

A value of 100 is a good trade-off between accuracy and runtime.

```
qbins: 100
```

Seed

This seed is useful for estimation reproducibility, in case when `dopseudohaploid: YES` (since this leads to randomly sample one of the alleles of heterozygous genotypes) and/or to select random individuals to create an outgroup population when `outpop: RANDOM`.

```
seed: 31
```

Name of a file containing the size of the chromosomes

Providing chromosome sizes is required for the chromosomal jackknife procedure, to calculate standard errors for the estimates. If the option is set as *None*, then the chromosome sizes will be inferred from the number of markers in each chromosome as read from the *.snp input file.

```
blocksizename: None
```

Resulting parameter file

We thus generate a parameter file that we will call *test.par* and which contains the following options:

```
genotypename: example.geno.gz
snpname: example.snp.gz
indivname: example.ind
targetpop: Pop1
outpop: None
outputprefix: results/test
chrom: 1, 2, 3, 4, 5
binsize: 0.001
mindis: 0.001
maxdis: 0.3
maxpropsharingmissing: 1
minmaf: 0
haploid: NO
dopseudohaploid: YES
morgans: NO
onlyfit: NO
usefft: YES
qbins: 100
seed: 31
blocksizename: None
```

Run *ASCEND*

To run the estimation, we simply need to run the command:

```
python3 ASCEND.py -p test.par
```

Output

Running *ASCEND* will generate all output files in the *results/* directory, with prefix *test.**. The files of major interest are:

- *test.png*
  This is a visual representation of the allele sharing decay curve. The blue points are the empirical values of the allele sharing correlation at each bin of genetic position (in cM). The red line is the single-factor exponential fit. Carefully check that the curve has decayed to 0, otherwise consider increasing `maxdis`. In the top-right corner, we report the estimated founder age (T_f) in generations before present, with the confidence interval at 95%; as well as the estimated founder intensity (I_f), reported in percents; and the NRMSD which is a measure of noisiness of the decay curve and a measure of fit quality. **The table of numerical values to re-plot the decay curve elsewhere can be found in *test.out* .**

- *test.est*
  This is a text file with the various estimates. Excluding the header, the first line reports (in order) the jackknife mean estimate, the standard error, the lower bound of the confidence interval at 95%, the upper bound of the confidence interval at 95% of the founder age. Second line for founder intensity (in %). Third line for the NRMSD.