

# Notes on ConStruct 1.0: An R Script to distinguish between substructure and consanguinity within a population using multilocus microsatellite data.

Andrew D.J. Overall

School of Pharmacy & Biomolecular Sciences, University of Brighton, BN2 4GJ, UK

## 1 Instructions

CON-STRUCT 1.0 is an R script that uses the method outlined in Overall & Nichols (2001) for partitioning the probable causes of excess homozygosity into that due to the Wahlund effect and that due to consanguinity. The details of the method can be found in Overall (2015) and an example input file is provided (`infile.txt`). Once the script has been copied into the working directory of R, three functions can be called:

1) `max.likelihood` - Estimates the magnitude of excess homozygosity ( $F$ ) within an existing dataset.

```
max.likelihood = function(data, max.alleles, resolution)
```

Arguments:

`data` is the input file of multilocus genotypes

`max.alleles` places an uppermost limit on the number of alleles considered

`resolution` is the resolution of the  $F$  parameter (i.e., the number of estimates made between 0 and the maximum value of  $F$ )

Example of use:

```
> max.likelihood(data="infile.txt", max.alleles=1000, resolution=100)
```

2) **construct** - Estimates the joint likelihood of the % of the population with consanguineous parents and  $F_{ST}$  within an existing dataset

```
construct = function(data, max.alleles, f.resolution, c.resolution, r)
```

Arguments:

**data** is the input file of multilocus genotypes

**max.alleles** places an uppermost limit on the number of alleles considered

**f.resolution** is the resolution of the Fst parameter

**c.resolution** is the resolution on the c parameter (% of population that is inbred)

**r** is the value of the inbreeding coefficient being considered for the analysis of the dataset

Example of use:

```
> construct(data="infile.txt", max.alleles=1000, f.resolution=100, c.resolution=100,
r=0.0625)
```

3) **simulate** - Simulates dataset with specified % of consanguinity and  $F_{ST}$  between two subpopulations

```
simulate = function(N, num.loc, fst, r.actual, c, r.consider, max.alleles, f.resolution,
c.resolution, iteration)
```

Arguments:

**N** is the total sample size

**num.loc** is the number of loci

**fst** is the value of Fst that is to be simulated between two populations

**r.actual** is the inbreeding coefficient of the inbred individuals

**c** is the proportion of the population inbred to degree r.actual

**r.consider** is the value of the inbreeding coefficient being considered for the analysis of the simulated dataset

**max.alleles** places an uppermost limit on the number of alleles considered

**f.resolution** is the resolution of the Fst parameter

**c.resolution** is the resolution on the c parameter

**iteration** is the number of iterations of the simulation run in order to arrive at the specified, simulated Fst

Example of use:

```
> simulate(N=200, num.loc=12, fst=0.05, r.actual=0.05, c=0.5, r.consider=0.05,
max.alleles=100, f.resolution=100, c.resolution=100, iteration=10000)
```

If the number of loci specified is, as in this example, 12, the code needs to be modified to tell it how many alleles are required for each locus, for example:

```
num.alleles = c(4,5,6,7,8,9,10,10,11,9,8,4)
```

## 2 References

Overall ADJ and Nichols RA. A method for distinguishing consanguinity and 204 population substructure using multilocus genotype data. *Mol. Biol. Evol.*, 18(11): 2048-2056, 2001.

Overall ADJ. CON-STRUCT 1.0: an R Script to distinguish between substructure and consanguinity within a population using multilocus microsatellite data. *Cogent Biology* (in review)