Notes on CON-STRUCT 1.0: R Scripts 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 Introduction & Instructions

CON-STRUCT 1.0 is a set of three R scripts that use 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 used in the scripts can be found in Overall (2015). Three scripts are available and should each be copied into the working directory of `R`. A brief outline of their implementation is provided here.

1.1 Option 1 - Measure excess homozygosity in your dataset

The first script, `ConStruct.Option.1.r`, identifies the overall excess in homozygosity over Hardy-Weinberg expectations, based upon the allele frequencies in the dataset. The maximum likelihood estimate of F (excess homozygosity) is output and a plot generated, along with the maximum value of F_{ST} for two subpopulations, according to Hedrick (2005). This script is run by typing

```
> source("ConStruct.Option.1.r").
```

1.2 Option 2 - Analyse existing dataset

The second script `ConStruct.Option.2.r` implements the method outlined in Overall & Nichols (2001) and Overall (2015) and estimates the joint maximum likelihood distribution for c_g (the proportion of the population that is inbred through consanguinity) and F_{ST} between unknown population substructure (the Wahlund effect). The maximum likelihood values are output, along with a contour plot of the distribution. In addition, the exp(likelihood) values are placed into an output file: `ConStruct.Outfile.txt`. This script is run by typing

`> source("ConStruct.Option.2.r").`

The user is then prompted for the input file. The format for the input file is given by the example dataset provided: "input.txt", which is a tab delimited series of multilocus diploid genotypes. Each line represents a different individual and missing genotypes are presented as `0 0`. The default limit in allele size (`max.alleles`) is 1000. Once the input file has been selected, you will be prompted to

`"Type in the value of consanguinity you wish to consider (eg, 0.125; 0.0625 etc)"`.

For example, if it is known that the population is unlikely to inbreed closer than first-cousins, then $R_g = 0.06$. The maximum likelihood estimate for the corresponding c_g is then an estimate of the most likely proportion of the population whose parents were related as first cousins. It is helpful at this stage to consider the output from `ConStruct.Option.1.r`. For example, if the maximum likelihood value of F was 0.06, then this is consistent with $F_{ST} = 0.06$, as well as $c_g = 1.00$ (when $R_g = 0.06$) and the two cannot be distinguished. Consanguinity and the Wahlund effect can only be disentangled when $R_g >$ F. So, if we type 0.125, then an F = 0.06 is consistent with $F_{ST} = 0.06$, as well as $c_g = 0.48$ (when $R_g = 0.125$), and the two can be distinguished.

1.3 Option 3 - Simulate dataset

It can be helpful to identify whether a given dataset has sufficient information for consanguinity and cryptic substructure to be disentangled. Option 3 generates simulated population data for a specified sample size, number of loci, number

of alleles for each locus, proportion of population that are consanguines and the magnitude of population substructure (F_{ST}) between two subpopulations. Because of the simulation involved, this script can take some time, depending on the magnitude of parameters. This script is run by typing

```
> source("ConStruct.Option.3.r").
```

For the simulated dataset, the user is then prompted for

```
Population Size (N):
Number of Loci:
Required value of F_ST:
Required degree of Consanguinity between Parents (eg, 0.25; 0.125 etc):
Required Proportion of Population that is Consanguineous (eg, 0.1; 0.5 etc):
```

This dataset is then analysed using the equivalent method of Option 2. The user is prompted for

```
Value of consanguinity being investigated (eg, 0.25; 0.125 etc):
```

This, of course, doesn't have to be the same as that you have simulated. You may, for example, be interested in exploring the sensitivity of the method when the incorrect value of consanguinity is assumed. The user is then prompted to type in the number of alleles for each locus:

```
Type Number of Alleles for each Locus:

[1] For Locus

[1] 1
```

As with Option 2, the maximum likelihood values are output, along with a contour plot of the distribution. Also, the exp(likelihood) values are placed into an output file: `ConStruct.Sim.Outfile.txt`.

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)

Hedrick PW. A standardized genetic differentiation measure. Evolution, 59:1633 -1638, 2005.