October 2007

TempoFs

This note documents the usage of the `TempoFs` program for estimating genetic drift and effective population size under the temporal method. `TempoFs` implements Jorde & Ryman's (2007) estimator, F_s, which should provide unbiased estimates also for highly polymorphic loci with many alleles in low frequency, including microsatellites. Both sample plans 1 and 2 (*sensu* Nei & Tajima 1981; Waples 1989) are supported.

Installing TempoFs

The downloaded zip file contains the executable program (TempoFs.exe for Windows/Dos and TempoFs for Linux) and source code (TempoFs.c).

Input data

Allele frequency data must be given blank-separated one locus at a time, and with samples in chronological order for each locus, and should be entered into a file formatted as follows.

Line 1: number of samples (integer), number of loci (integer), and optionally: any text

Line 2: number of alleles at each locus (integer)

Line 3 - onwards: allele frequency (real) for the first locus, number of samples genes (2x number of diploid individuals; real), name of locus (text) and generation (real).

Line 3 is repeated for this locus for each additional sample

Lines 3+samples are repeated for each additional locus, keeping sample order as before

Example data file:

```
3 9  Data are from Begon, Krimbas & Loukas (1980): Heredity, 45: 335-350.
3 6 3 4 3 4 3 5 5
0.289 0.674 0.037                          380.0  EST3 1
0.325 0.621 0.054                          500.0  EST3 8
0.255 0.689 0.056                          670.0  EST3 10
0.000 0.026 0.405 0.527 0.021 0.021        380.0  EST5 1
0.016 0.068 0.300 0.523 0.093 0.000        500.0  EST5 8
0.011 0.057 0.374 0.503 0.049 0.006        670.0  EST5 10
0.058 0.010 0.932                          380.0  EST7 1
0.029 0.003 0.968                          500.0  EST7 8
0.047 0.000 0.953                          670.0  EST7 10
0.000 0.453 0.547 0.000                    380.0  APH  1
0.010 0.568 0.412 0.010                    500.0  APH  8
```

```
0.012 0.532 0.442 0.015                  670.0  APH  10
0.042 0.921 0.037                        380.0  PGM  1
0.055 0.930 0.015                        500.0  PGM  8
0.033 0.940 0.056                        670.0  PGM  10
0.005 0.400 0.492 0.103                  380.0  HK1  1
0.003 0.385 0.509 0.103                  500.0  HK1  8
0.003 0.405 0.473 0.119                  670.0  HK1  10
0.016 0.947 0.037                        380.0  ME   1
0.015 0.941 0.041                        500.0  ME   8
0.013 0.924 0.062                        670.0  ME   10
0.021 0.142 0.674 0.116 0.047            380.0  XDH  1
0.012 0.097 0.768 0.091 0.032            500.0  XDH  8
0.020 0.118 0.724 0.092 0.046            670.0  XDH  10
0.116 0.237 0.547 0.053 0.047            380.0  AO   1
0.080 0.188 0.600 0.088 0.047            500.0  AO   8
0.093 0.158 0.599 0.092 0.059            670.0  AO   10
```

A correctly formatted input file may be generated from individual genotype data in GENEPOP format using the included GPtoSU utility.

Running TempoFs

The program (`TempoFs` or `TempoFs.exe`) takes the name of the input file on the command line:

```
TempoFs mydata.txt
```

or, if no file name is given, ask for it during execution. In the latter case the program also ask for the name of an output file to write the results (otherwise the name of the output file defaults to <input.name>.OUT):

```
Enter name of data file: mydata.txt
Enter output file name: mydata.result
```

Finally, the program asks for which sample plan was used [1 or 2: see Nei & Tajima (1981) and Waples (1989) for details]. In the case of plan 1, the actual number of individuals in the population is also requested:

```
Enter sample plan (1 or 2): 1
Enter actual (census) population size, N: 100
```

Note that the program will generate an error during execution if allele frequencies do not add to unity at any locus (+/- a small quantity `AlleleFreqTolerance` allowing for minor rounding errors). Such an error will be generated for the example data above:

```
Error in indata file: allele frequencies sum to 1.029000 at locus 5 in
sample 3
```

In such circumstances it may be possible to adjust `AlleleFreqTolerance` in the source code upwards to accommodate reasonable rounding errors and

recompile the program. However, in this particular case there is probably a typographical error in the original paper and it would probably be best to try to weed out errors in the data before attempting to estimate anything from them.

Output

The program estimates genetic drift, `F_s'` (corrected for sampling according to the appropriate sample plan), and effective population size per generation, `Ne`, assuming that time (the last column in the input file) is given in generations. Estimates are calculated as described in Jorde & Ryman (2007), separately for each consecutive sample interval in the case of more than two samples.

The program reports `Fs` (uncorrected), `Fs'` (corrected for sampling according to the appropriate sampling plan), and `Ne` (per generation) separately for each locus and over all loci, for each sample interval. Also reported are the number of alleles at each locus (`K`), the sample sizes (`S1` and `S2`), the number of generations (`t`) between the two samples, and descriptive notes. Example output:

```
Output from TempoFs (version Oct. 22. 2007)
Data are from from file begon_etal.dat

Note: Effective size (Ne = t/2Fs') is estimated under the assumptions that:
         - the organism is diploid
         - generations are discrete
         - time is measured in generations
         - only genetic drift affects allele frequency dynamics
         - samples are from a single population, drawn according to plan 2
         - multiple sample intervals are computed separately

Estimated drift (Fs') and Ne over the interval from    1 to    8:
     Locus     K      S1      S2     Fs          Fs'          t          Ne
      EST3     3     190     250     0.0090706    0.0044273    7.0        791
      EST5     6     190     250     0.0314322    0.0266087    7.0        132
      EST7     3     190     250     0.0228744    0.0181490    7.0        193
       APH     4     190     250     0.0621295    0.0566610    7.0         62
       PGM     3     190     250     0.0052271    0.0005899    7.0       5933
       HK1     4     190     250     0.0008857   -0.0037535    7.0       -932
        ME     3     190     250     0.0005019   -0.0041380    7.0       -846
       XDH     5     190     250     0.0260117    0.0212546    7.0        165
        AO     5     190     250     0.0126601    0.0080046    7.0        437
  All loci    36     190.0   250.0   0.0217239    0.0170090    7.0        206
Jackknife over 9 loci (assumed independent):
                     Mean:          0.0217174    0.0170202    7.0        206
                       SE:          0.0081071    0.0080513
95% Confidence intervals (Mean -/+ 1.96*SE):
       Lower ( 2.5%) limit:          0.0058274    0.0012397              107
       Upper (97.5%) limit:          0.0376074    0.0328008             2823
```

```
Estimated drift (Fs') and Ne over the interval from    8 to   10:
     Locus     K      S1      S2     Fs           Fs'          t          Ne
      EST3     3     250     335     0.0196921    0.0161271    2.0         62
      EST5     6     250     335     0.0130169    0.0094963    2.0        105
      EST7     3     250     335     0.0073427    0.0038425    2.0        260
       APH     4     250     335     0.0043171    0.0008211    2.0       1218
       PGM     3     250     335     0.0165233    0.0129822    2.0         77
       HK1     4     250     335     0.0033053   -0.0001902    2.0      -5257
        ME     3     250     335     0.0058442    0.0023466    2.0        426
       XDH     5     250     335     0.0062545    0.0027563    2.0        363
        AO     5     250     335     0.0020697   -0.0014260    2.0       -701
  All loci    36     250.0   335.0   0.0081840    0.0046817    2.0        214
Jackknife over 9 loci (assumed independent):
                      Mean:          0.0081908    0.0046925    2.0        213
                        SE:          0.0025569    0.0025529
95% Confidence intervals (Mean -/+ 1.96*SE):
       Lower ( 2.5%) limit:          0.0031793   -0.0003112               103
       Upper (97.5%) limit:          0.0132024    0.0096961              -3213

Where
K = number of alleles
S1, S2 = number of genotyped individuals in each of the two samples
Fs = observed allele frequency shift over the sample interval
Fs' = d.o. corrected for expected contribution from sampling
t = number of generations between the two samples
Ne = estimated variance-effective population size per generation
```

Note that negative estimates (F_s' and N_e) imply that the observed temporal allele frequency shift was less than that expected from random sampling errors alone, and such estimates should be interpreted as lack of evidence for genetic drift, implying an infinite estimate of N_e. Some negative estimates are unavoidable in any unbiased estimator when the parameter is close to zero.

The `TempoFs` program uses the standard delete-one jackknife over loci to evaluate uncertainty in the estimates (as suggested by Weir & Cockerham 1984). Reported are the mean (`Mean`) and standard error (`SE`) of the jackknife estimates of F_s and F_s'. `Mean` should be close to the original `All loci` estimate. The jackknife Mean and SE are used to put a 95% confidence interval (CI) for mean F_s and F_s', assuming that they are normally distributed. A 95% CI for the estimated N_e is then is calculated from the lower and upper limits for F_s'. In the example above, the point estimates for N_e is 206 per generation during the first interval (generation 1 to 8) and 213 for the second interval (generation 8 to 10). The corresponding 95% CIs are very wide, however, ranging from 107 to 2823 and from 103 to infinity, respectively.

Given the similarity of point estimates for the two intervals it may be reasonable to assume that N_e remains constant over time and to combine the two into a single estimate over the 9 generations. One way of combining data

for the two intervals are to weight each interval by the number of generations between the two samples. The rationale for such a weighing scheme is that drift accumulates over generations and the longer interval should be more informative than the shorter one.

First, we calculate average amount of drift per generation for the two intervals: Fs'(1-8) = 0.0170202/7 = 0.002431457 and Fs'(8-10) = 0.0046925/2 = 0.00234625. The weighted mean (per generation) is Fs'(1-10) = (7/9)*Fs'(1-8) + (2/9)*Fs'(8-10) = 0.0170202 + 0.0046925 = 0.0024125. An estimate for effective size over the entire period then becomes Ne = 1/(2*Fs'(1-10)) = 1/(2*0.0024125) = 207.3.

The standard error of the combined estimate may be calculated from the variance of a weighted sum of variances (cf. Sokal & Rohlf 1981, p. 135, equation 7.1), making the (questionable) assumption that the two estimates are independent. The variance for the average F_s' per generation for each of the two periods is: Var(1-8) = $(0.0080513/7)^2 = 1.322927e^{-6}$, and Var(8-10) = $(0.0025529/2)^2 = 1.629325e^{-6}$. Thus, the variance for the entire period is Var(1-10) = $(7/9)^2 * 1.322927e^{-6} + (2/9)^2 * 1.629325e^{-6} = 8.807498e^{-7}$, and the standard error is the square root of this variance: SE(1-10) = Var(1-10)$^{1/2}$ = 0.00093848. Using the normal assumption as before yields the 95% CI for the average F_s' per generation: Fs'(1-10) -/+ 1.96*SE(1-10), or from 0.0005731 to 0.0042519. The 95% CI for N_e over the period is therefore 1/(2*0.0042519) to 1/(2*0.0005731), or from 117.6 to 872.4. The CI is still quite wide, but considerably improved relative to the two separate estimates, particularly for the upper limit.

Citation

To cite this program refer to Jorde & Ryman (2007).

Per Erik Jorde
University of Oslo
Department of Biology
Centre for Ecological and Evolutionary Synthesis
P.O. Box 1066 Blindern
N-0316 Oslo
Norway
E-mail: p.e.jorde@bio.uio.no

References

Begon, M., C.B. Krimbas, & M. Loukas, 1980. The genetics of *Drosophila subobscura* populations. XV. The effective size of a natural population estimated by three independent methods. Heredity **43:** 335-350.

Jorde, P.E., & N. Ryman, 2007. Unbiased estimator for genetic drift and effective population size. Genetics **177:** 927-935.

Nei, M. & F. Tajima, 1981. Genetic drift and estimation of effective population size. Genetics **98:** 625-640.

Weir, B.S., & C.C. Cockerham, 1984. Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358-1370.

Sokal, R.S., & F.J. Rohlf, 1981. *Biometry.* 2nd ed. Freeman, New York, NY.

Waples, R.S., 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121:** 379-391.