[1] Computation of minimum spanning network between haplotypes is only possible if a distance matrix is provided or if it can be computed from the data.

# *8.1 Intra-population level methods*

## 8.1.1 Standard diversity indices

### *8.1.1.1 Gene diversity*

Equivalent to the expected heterozygosity for diploid data. It is defined as the probability that two randomly chosen haplotypes are different in the sample. Gene diversity and its sampling variance are estimated as

$$\hat{H} = \frac{n}{n-1}(1-\sum_{i=1}^{k} p_i^2)$$

$$\mathrm{V}(\hat{H}) = \frac{2}{n(n-1)}\left\{2(n-2)\left[\sum_{i=1}^{k} p_i^3 - (\sum_{i=1}^{k} p_i^2)^2\right] + \sum_{i=1}^{k} p_i^2 - (\sum_{i=1}^{k} p_i^2)^2\right\},$$

where *n* is the number of gene copies in the sample, *k* is the number of haplotypes, and $p_i$ is the sample frequency of the *i*-th haplotype.

Note that Arlequin outputs the standard deviation of the Heterozygosity computed as $s.d.(\hat{H}) = \sqrt{V(\hat{H})}$.

*Reference*:
Nei, 1987, p.180.

### *8.1.1.2 Expected heterozygosity per locus*

For each locus, Arlequin provides an estimation of the expected heterozygosity simply as

$$\hat{H} = \frac{n}{n-1}(1-\sum_{i=1}^{k} p_i^2)$$

### *8.1.1.3 Number of usable loci*

Number of loci that show less than a specified amount of missing data. The maximum amount of missing data must be specified in the *General Settings* tab dialog .

### *8.1.1.4 Number of polymorphic sites (S)*

Number of usable loci that show more than one allele per locus.

### *8.1.1.5 Allelic range (R)*

For MICROSAT data, it is the difference between the maximum and the minimum number of repeats.

#### *8.1.1.6 Garza-Williamson index (G-W)*

Following Garza and Wlliamson (2001), the G-W statistic is given as $G-W=\frac{k}{R+1}$ where *k* is the number of alleles at a given loci in a population sample, and R is the allelic range. Originally, the denominator was defined as just *R* in Garza and Wlliamson (2001), but this could lead to a division by zero if a sample is monomorphic. This adjustment was introduced in Excoffier et al. (2005).

This statistic was shown to be sensitive to population bottleneck, because the number of alleles is usually more reduced than the range by a recent reduction in population size, such that the distribution of allele length will show "vacant" positions. Therefore the G-W statistic is supposed to be very small in population having been through a bottleneck and close to one in stationary populations.

Here we just report the statistics but do not provide any test.

### 8.1.2 Site Frequency Spectrum

The Site Frequency Spectrum (SFS) can be used for demographic inference. The user is referred to the fastsimcoal2 manual, available on http://cmpg.unibe.ch/software/fastsimcoal2/ for all relevant information on the structure of the generated files and on their use for parameter inference.

### 8.1.3 Molecular indices

#### *8.1.3.1 Mean number of pairwise differences ($\pi$)*

Mean number of differences between all pairs of haplotypes in the sample. It is given by

$$\hat{\pi}=\frac{n}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{k}p_i p_j \hat{d}_{ij},$$

where $\hat{d}_{ij}$ is an estimate of the number of mutations having occurred since the divergence of haplotypes *i* and *j*, *k* is the number of haplotypes, $p_i$ is the frequency of haplotype *I*, and *n* is the sample size. The total variance (over the stochastic and the sampling process), assuming no recombination between sites and selective neutrality, is obtained as

$$\mathrm{V}(\hat{\pi})=\frac{3n(n+1)\hat{\pi}+2(n^2+n+3)\hat{\pi}^2}{11(n^2-7n+6)}. \qquad \text{(Tajima, 1993)}$$

Note that similar formulas are also used for *Microsat* and *Standard* data, even though the underlying assumptions of the model may be violated. Note also that Arlequin outputs the standard deviation computed as $s.d.(\hat{\pi})=\sqrt{V(\hat{\pi})}$.

*References*:
Tajima, 1983
Tajima, 1993

### *8.1.3.2 Nucleotide diversity or average gene diversity over L loci*

It is computed here as the probability that two randomly chosen homologous (nucleotide or RFLP) sites are different. It is equivalent to the gene diversity at the nucleotide level for DNA data.

$$\hat{\pi}_n = \frac{\sum_{i=1}^{k} \sum_{j<i} p_i p_j \hat{d}_{ij}}{L}$$

$$\mathrm{V}(\hat{\pi}_n) = \frac{n+1}{3(n-1)L}\hat{\pi}_n + \frac{2(n^2+n+3)}{9n(n-1)}\hat{\pi}_n^2$$

Note that similar formulas are used for computing the average gene diversity over *L* loci for Microsat and Standard data, assuming no recombination and selective neutrality. As above, one should be aware that these assumptions may not hold for these data types. Note also that Arlequin outputs the standard deviation computed as $s.d.(\hat{\pi}_n) = \sqrt{V(\hat{\pi}_n)}$ .

Note that for RFLP data this measure should be considered as the average heterozygosity per RFLP site, which is different from the true diversity at the nucleotide level, for which one would need to know the base composition of the restriction sites.

*References*:
Tajima, 1983
Nei, 1987, p. 257

### *8.1.3.3 Theta estimators*

Several methods are used to estimate the population parameter $\theta = 2Mu$ , where *M* is equal to 2 *N* for diploid populations of size *N* , or equal to *N* for haploid populations, and *u* is the overall mutation rate at the haplotype level.

#### 8.1.3.3.1 Theta(Hom)

The expected homozygosity in a population at equilibrium between drift and mutation is usually given by

$$H = \frac{1}{\theta + 1}.$$

However, Zouros (1979) has shown that this estimator was an overestimate when estimated from a single or a few loci. Although he gave no closed form solution, Chakraborty and Weiss (1991) proposed to iteratively solve the following relationship between the expectation of $\hat{\theta}_H$ and the unknown parameter $\theta$