

# VALUES AND STATISTICS INDICATED IN MSTATSPOP:

---

## Effective length for each population:

**Eff\_length1\_pop\_outg[0]:** Effective length1 for population 0. That is, considering the outgroup, how many positions have at least one sequence per population 0 and the outgroup exist.

Useful in case considering missing positions (-u 1). In case -u 0 all lengths are equal. Useful for calculating divergence vs outgroup or versus populations.

**Eff\_length2\_pop\_outg[0]:** Effective length2 for population 0. That is, considering the outgroup, how many positions have at least two sequences per population 0 and the outgroup exist.

Useful in case considering missing positions (-u 1). In case -u 0 all lengths are equal. Useful for calculating levels of variability.

**Eff\_length3\_pop\_outg[0]:** Effective length3 for population 0. That is, considering the outgroup, how many positions have at least three sequences per population 0 and the outgroup exist.

Useful in case considering missing positions (-u 1). In case -u 0 all lengths are equal. Useful for calculating tests of neutrality (variances).

**Eff\_length1\_pop[0]:** Effective length1 for population 0. That is, how many positions have at least one sequence per population 0.

Useful in case considering missing positions (-u 1). In case -u 0 all lengths are equal. Useful for calculating divergence versus populations.

**Eff\_length2\_pop[0]:** Effective length2 for population 0. That is, how many positions have at least two sequences per population 0.

Useful in case considering missing positions (-u 1). In case -u 0 all lengths are equal. Useful for calculating levels of variability.

**Eff\_length3\_pop[0]:** Effective length3 for population 0. That is, how many positions have at least three sequences per population 0.

Useful in case considering missing positions (-u 1). In case -u 0 all lengths are equal. Useful for calculating tests of neutrality (variances).

In principle, if the analysis contains an outgroup, all analysis are conditioned to the presence of the outgroup variant.

## STATISTICS:

Estimates of variability for each population:

**S[0]:** Number of segregating sites at population 0

**Theta(Wat)[0]:** Watterson (1975) estimation of variability.

**Theta(Taj)[0]:** Tajima (1983) estimation of variability.

**Theta(Fu&Li)[0]:** Fu and Li's (1991) estimation of variability.

**Theta(Fay&Wu)[0]:** Fay and Wu's (2000) estimation of variability.

**Theta(Zeng)[0]:** Zeng's (2006) estimation of variability.

**Theta(Achaz,Wat)[0]:** Watterson estimator without considering singletons (Achaz 2008)

**Theta(Achaz,Taj)[0]:** Tajima estimator without considering singletons (Achaz 2008)

**Divergence[0]:** Divergence between population 0 and the outgroup.

**an\_x[0]:**  $(\sum_{i=1}^{n-1} 1/i)$ , that is, sumatory of  $(1/i)$  from  $i=1$  to  $n-1$ , being  $n$  the number of samples. In case missing data (-u 1), sum of an for each position and divided by the eff\_length2\_pop. Used for Watterson estimator and neutrality tests.

**bn\_x[0]:**  $(\sum_{i=1}^{n-1} 1/i^2)$ , that is, sumatory of  $(1/i^2)$  from  $i=1$  to  $n-1$ , being  $n$  the number of samples. In case missing data (-u 1), sum of an for each position and divided by the eff\_length2\_pop. Used for neutrality tests.

**an\_xo[0]:** Same but considering only those positions were the outgroup is present.

**bn\_xo[0]:** Same but considering only those positions were the outgroup is present.

Haplotype diversity and number of haplotypes for each population:

**HapW[0]:** Haplotype diversity at population 0.

**nHap[0]:** number of haplotypes at population 0.

Neutrality tests for each population:

**Tajima D[0]:** Test of Tajima (1989)

**Fu&Li D[0]:** Test of Fu and Li D (1993)

**Fu&Li F[0]:** Test of Fu and Li D (1993)

**Fay&Wu norm H[0]:** Test of Fay and Wu (2000) normalized (Achaz 2009, Zeng 2006)

**Fay&WuH[0]:** Test of Fay and Wu (2000)

**Ferretti L[0]:** Test of Ferretti similar to Fay and Wu but using theta Watt instead Theta Taj (Ferretti et al.)

**Zeng E[0]:** Test of Zeng (2006)

**Achaz Y[0]:** Test of Achaz (2008)

**Fs[0]:** Test Fs (Fu 1997)

**R2[0]:** Test of R2 (2002)

Variants assigned to exclusive, fixed, polymorphic but fixed in rest of pops, and shared:

Ss[rest] are shared variants between populations but fixed within:

**Sx[0]:** Exclusive variants at population 0

**Sf[0]:** Fixed variants at population 0

**Ss:** Shared variants among populations

**Sxf[0]:** Exclusive variants at population 0 that are fixed at other populations.

Mismatch distribution statistics:

**SDev[0]:** Standard deviation of Tajima's estimator.

**Skewness[0]:** Third standardised moment of Tajima's estimator.

**Kurtosis[0]:** Fourth standardised moment of Tajima's estimator.

Differentiation statistics:

**Differentiation with nucleotide sequences (even with missing data):**

**Fstall:**  $1 - \frac{\text{mean}(\pi_{\text{within}}/\text{nt})}{\text{mean}(\pi_{\text{among}}/\text{nt})}$

**Fst1all[pop1]:**  $1 - \frac{\text{mean}(\pi_{\text{within}}/\text{nt}[\text{rest}], \pi_{\text{within\_rest}}/\text{nt}[\text{pop1}])}{\text{mean}(\pi_{\text{among}}/\text{nt})}$

**Fst[pop1][pop2]:**  $1 - \frac{\text{mean}(\pi_{\text{within}}/\text{nt}[\text{pop1}], \pi_{\text{within}}/\text{nt}[\text{pop2}])}{\pi_{\text{among}}/\text{nt}[\text{pop1}][\text{pop2}]}$  (Hudson et al. 1992)

### Differentiation with haplotype data (phase is necessary):

**Fsthall:**  $1 - \frac{\text{mean}(\pi_{\text{h\_within}})}{\text{mean}(\pi_{\text{h\_among}})}$

**Fsth1all[pop1]:**  $1 - \frac{\text{mean}(\pi_{\text{h\_within}}[\text{pop1}] - \pi_{\text{h\_within}}[\text{rest}])}{\text{mean}(\pi_{\text{h\_among}})}$

**Fsth[pop1][pop2]:**  $1 - \frac{\text{mean}(\pi_{\text{h\_within}}[\text{pop1}], \pi_{\text{h\_within}}[\text{pop2}])}{\pi_{\text{h\_among}}[\text{pop1}][\text{pop2}]}$

Frequency of variants for each population:

**fr[0,1]:** Site frequency Spectrum per population

Frequency of each haplotype in the populations:

**frH[0,hap00]:** frequency of each haplotype per population

Joint frequency distribution for each variant and population (No included variants that are missing or polymorphic in the outgroup):

**SNP[xx]:** frequency of each SNP at each population.

Relative Site Frequency Spectrum

**rSFS[pop,freq]:** number of variants at frequency freq in the population pop, relative to the whole sample (excluding outgroup, only used for polarizing)