

## Conceptual Foundations of Common Tests in Molecular Population Genetics

### Tajima's D

Recall that the expected time to coalescence for two alleles is  $E[t_c] = 2N$ . If mutations are occurring at rate  $\mu$  then each allele is expected to have acquired  $\mu E[t_c] = 2N\mu$  mutations during this time. That means the expected number of differences between the two alleles will be

$$E[I] = 2 * \mu t_c = 4N\mu = \theta$$

(In practice, this measure is often standardized by the length of the allele, i.e., number of sites  $n$ , to give  $\pi = I/n$ . This is the same as using the mutation rate per site ( $\mu_s$ ) rather than the mutation rate for the entire allele,  $\mu = n * \mu_s$ .)

Using the equation, above we could estimate  $\theta$  as from the average number of differences between all possible pairs in a sample of alleles,  $\theta_\pi = \bar{\Pi}$ . The subscript is to remind us that this is an estimate of  $\theta$  obtained using pairwise differences.

Also remember the expected total number of variable sites in sample of  $m$  alleles from a single species is given by

$$E(X) = 4N\mu k_m = \theta k_m \quad \text{where } k_m = \sum_{i=1}^{m-1} 1/i$$

(One could divide by the number of sites to obtain  $x = X/n$  to get the number of variants per site in a sample of  $m$  alleles. As above, this would be equivalent to using  $\mu_s$  in place of  $\mu$ .)

Using the equation, above we could estimate  $\theta$  as from the total number of polymorphic samples found in a sample of  $m$  alleles,  $\theta_W = X/k_m$ . This is known as Watterson's estimator of  $\theta$ ; the subscript reminds us of that.

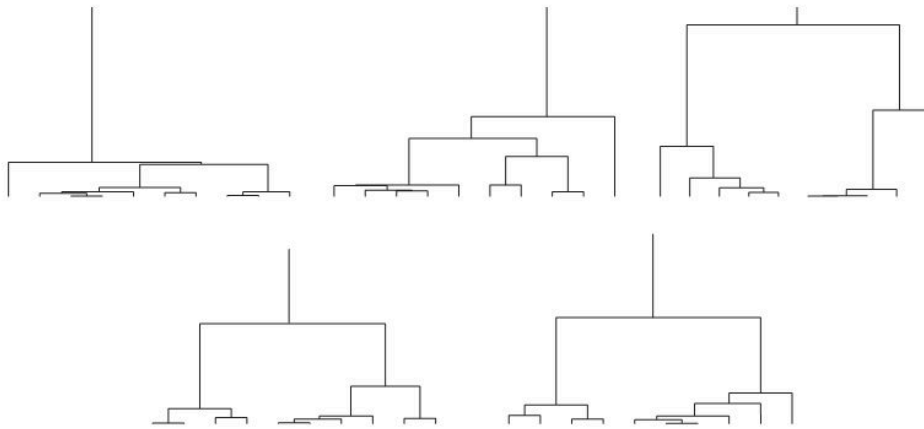
Tajima's D is a metric based on the difference between these two estimators of  $\theta$

$$\text{Tajima's } D = \frac{\theta_\pi - \theta_W}{\sqrt{\text{Var}(\theta_\pi - \theta_W)}}$$

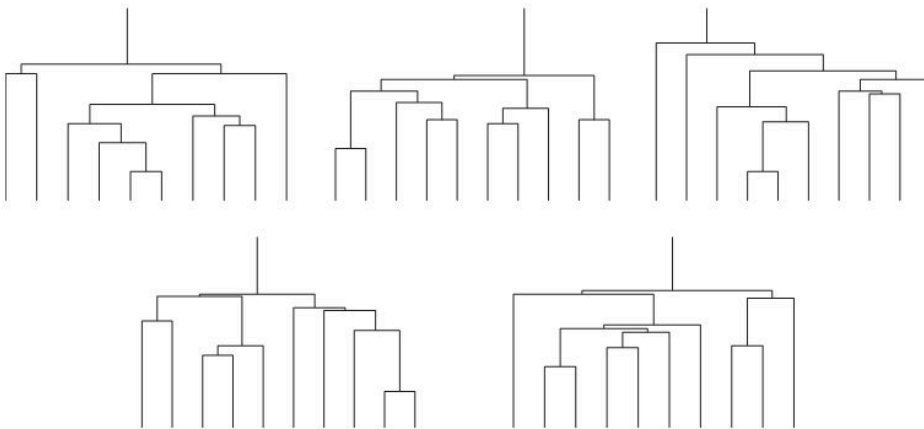
As both  $\theta_\pi$  and  $\theta_W$  are estimators of  $\theta$  they should be equal if the alleles in the sample are evolving under the assumptions we have considered, i.e., evolving neutrally and at mutation-drift equilibrium. If these assumptions are violated (alleles not evolving neutrally or out of equilibrium due to demographic changes), then the distribution of alleles in a population will not be as expected under the assumptions; consequently estimates based on  $\bar{\Pi}$  and  $X/k_m$  will not be a simple reflection of  $4N\mu$ . The two estimators  $\theta_\pi$  and  $\theta_W$  are affected differently by different kinds of deviations in the frequency distribution of alleles away from the neutral expectation.  $\theta_\pi$  is more sensitive to an excess of alleles at intermediate frequencies.  $\theta_W$  is more sensitive to an excess of alleles at low frequencies. Under neutrality  $D$  should be close to zero. When there is an

excess of alleles at intermediate frequencies (as expected under balancing selection or recent reductions in population size),  $D$  will be positive. When there is an excess of alleles at low frequencies (as expected under purifying selection or recent expansions in population size),  $D$  will be negative.

Demographic effects (changes in population size, population structure) affect the shape of genealogies and thus can affect  $D$ .



**Figure 4.2** Five replicates of the coalescent process with constant population size for a sample of ten genes. Note the large variance in the time of the MRCA among replicates.



**Figure 4.3** Five replicates of the coalescent with exponential growth,  $\beta = 1000$ , for a sample of  $n = 10$  genes. Note the smaller variance in the time until the MRCA compared to the same quantity in Figure 4.2.

Copied from Hein, Schierup, and Wiuf. 2004. *Gene Genealogies, Variation and Evolution*.

As illustrated above, exponentially growing populations tend to have an excess of long terminal branches, which will result in abundance of rare variants, making *Tajima's D*  $< 0$ .

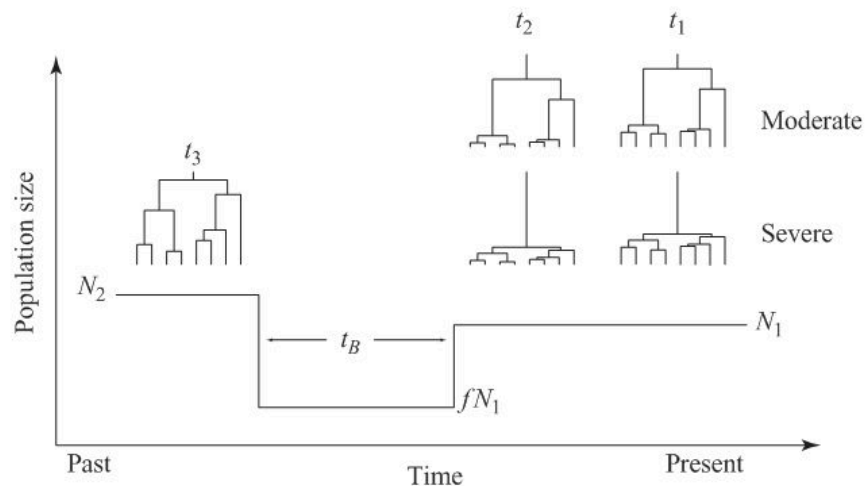
Bottlenecks are expected to reduce the total length of the tree but the effects of bottlenecks on the shape of the genealogy are somewhat more complicated because it depends on how long ago the bottleneck occurred and how severe it was. If a bottleneck occurred in the very distant past, it is unlikely to have any effect on the genealogy.

\*\*\*\*\*

If a severe bottleneck ended  $g$  generations ago and the population has since been at size  $N_1$ , how long ago would  $g$  have to be for you to be relatively certain there would be no effect on the genealogy of current day samples. Explain your answer.

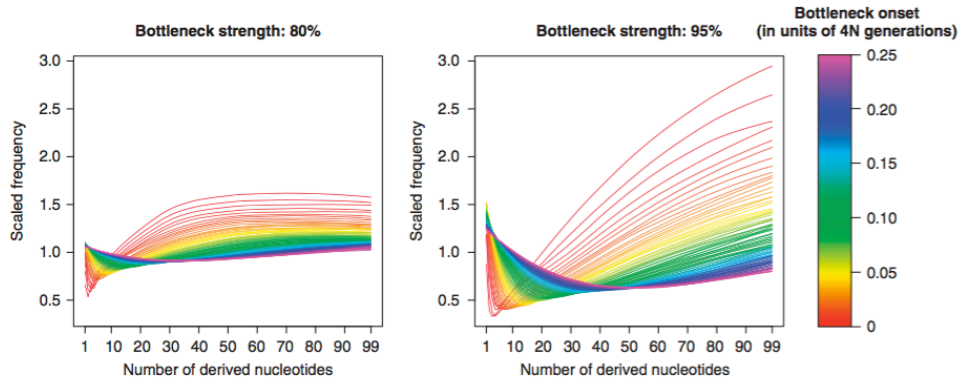
\*\*\*\*\*

If a bottleneck occurred in the less distant (but not too recent) past, then most current day lineages will coalesce during the bottleneck and there will have been very few coalescences between the present time (when  $N$  was large) and the bottleneck. Consequently, the genealogy will tend to have long external branches, resulting in an excess of rare variants (*Tajima's*  $D < 0$ ). For very recent bottleneck there can be an excess of intermediate frequency variants (*Tajima's*  $D > 0$ ) reflecting the divergence of the few lineages that did *not* coalesce during the bottleneck but further back in time. (Though if the bottleneck is sufficiently strong, then all current lineages will coalesce during the bottleneck and *Tajima's*  $D < 0$ . This explains the difference between the autosomal and mitochondrial results given in Fig. 3 of Gattepaille *et al.* (2013) shown below.)

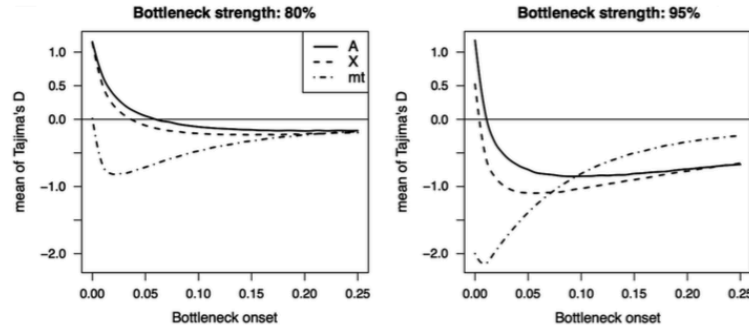


**Figure 4.8** The effect of a population bottleneck on the genealogical tree of a sample. The shape of the genealogical tree depends on when the genes are sampled (or equivalently when the bottleneck occurred relative to the time of sampling). In the figure example trees are shown when sampling occurs after the bottleneck, shortly before the bottleneck and some time before the bottleneck (time is running backwards).  $f$  denotes the fraction of the current population surviving through the bottleneck and  $t_B$  is the duration of the bottleneck. Severe bottleneck: most coalescence events occur during the bottleneck. Moderate bottleneck: some lineages are likely to survive through the bottleneck and find MRCA's while the population has size  $N_2$ .

Copied from Hein, Schierup, and Wiuf. 2004. *Gene Genealogies, Variation and Evolution*.



**Figure 2** Scaled SFS under a bottleneck model. The scaled SFS is the ratio of the SFS and the expected SFS for a model of constant population size. A sample of 100 chromosome of 50 000 bp long was simulated for a population of 10 000 present-day diploid individuals. (a) Recombination rate of  $1.5 \times 10^{-8}$  per site and per generation as well as a mutation rate of  $1.2 \times 10^{-8}$  per site and per generation was used (Scally and Durbin, 2012). Each curve corresponds to the ratio between the observed SFS under a bottleneck model and the expected SFS under a standard neutral model with constant population size. The color of the curves indicates the onset of the bottleneck, in units of four times the present-day population size. Each bottleneck lasted 1000 generations and the strength of the bottleneck measures the reduction in population size during the bottleneck: 80% reduction (left panel) or 95% reduction (right panel). A total of 20 000 simulations per model were performed with the software ms (Hudson, 2002).



**Figure 3** Mean of Tajima's  $D$  as a function of the time since the onset of the bottleneck. Average values of Tajima's  $D$  over 20 000 replicate simulations of bottleneck models, for an 80% reduction in population size (left panel) and a 95% reduction in population size (right panel). Different chromosomes are considered: autosome (A, solid line), X chromosome (X, dashed line) and mitochondrion (mt, dash-dotted line). The bottleneck lasted 1000 generations for all models, and the onset of the bottleneck varies, as indicated on the x axis (given in units of four times the present-day population size). The simulated samples are composed of 100 segments of 50 000 bp (the recombination rate is  $1.5 \times 10^{-8}$  per site and per generation), with a mutation rate of  $1.2 \times 10^{-8}$  per site and per generation, and they are randomly sampled from a present-day population of 10 000 diploid individuals. mtDNA segments are also 50 000 bp long, but they are non-recombining and the mutation rate is  $2.5 \times 10^{-6}$ . Time is measured in units of  $4N$  generations.

Copied from Gattepaille, Jakobsson, and Blum (2013, *Heredity*).

*Population subdivision* can also affect the shape of genealogies, but the effect depends on how the samples were collected over space. When multiple samples are collected from each of several demes, there will be an excess of long internal branches (due to longer coalescence times for between-deme samples), resulting in an excess of intermediate variants and *Tajima's D*  $> 0$ . If only a single sample is taken from each deme ("scattered sample") then there is little effect of subdivision.

### Tests involving both polymorphism and divergence.

Above we considered only polymorphism. Now consider the expected number of differences,  $E(D)$ , between two alleles that each come from a separate species that speciated  $G$  generations ago. The expected number of differences will depend on the expected time to coalescence of these two alleles. The expected time to coalescence for

these alleles will be  $E(T) = G + 2N_A$  because there are  $G$  generations while they are in separate species so it is impossible for them to coalesce and the expected time to coalescence for the two alleles once they are both in the same species is  $2N_A$  where  $N_A$  is the size of the ancestral species. Because mutations can accumulate along each of the two branches of the genealogy,

$$E(D) = 2\mu E(T) = 2\mu(G + 2N_A)$$

### McDonald-Kreitman Test

The purpose of this test is to determine whether a protein sequence is evolving neutrally. Changes in the DNA can be categorized as synonymous or non-synonymous changes. Synonymous changes are those that do not affect the amino acid sequence of the protein (e.g., third site in codons are degenerate). Such changes are assumed to be neutral in this test. Non-synonymous changes do affect the amino acid sequence and we want to know if we can reject the hypothesis that such changes evolve neutrally. We cannot simply compare the rate of evolution of synonymous and non-synonymous changes because such differences in the rate of evolution could be a result of differences in mutation rate, i.e.,  $\mu_{syn} \neq \mu_{non}$ . To test against neutrality, **we calculate the patterns expected if both synonymous and non-synonymous sites evolved neutrally**. In doing this test, you must have multiple alleles from one species and at least one allele from a related species.

Let us consider the number of sites polymorphic for synonymous sites in a sample of  $m$  alleles from our focal species,  $X_{syn}$ . We know that

$$E(X_{syn}) = 4N\mu_{syn}k_m$$

Similarly, the number of sites polymorphic for non-synonymous sites in this sample should be

$$E(X_{non}) = 4N\mu_{non}k_m$$

The expected number of differences at synonymous sites in a comparison between alleles from separate species is

$$E(D_{syn}) = 2\mu_{syn}(G + 2N_A)$$

The expected number of differences at nonsynonymous sites in a comparison between alleles from separate species is

$$E(D_{non}) = 2\mu_{non}(G + 2N_A)$$

Now let us consider the ratio for the expected amount of polymorphism to the expected amount of divergence for the two types of sites. For synonymous sites,

$$\frac{E(X_{syn})}{E(D_{syn})} = \frac{4N\mu_{syn}k_m}{2\mu_{syn}(G + 2N_A)} = \frac{2Nk_m}{G + 2N_A}$$

and for non-synonymous sites

$$\frac{E(X_{non})}{E(D_{non})} = \frac{2Nk_m}{G + 2N_A}$$

Therefore, **if both types of sites evolve neutrally**, we expect these ratios to be equal  $\frac{E(X_{syn})}{E(D_{syn})} = \frac{E(X_{non})}{E(D_{non})}$ . Sufficiently large deviations from equality indicate can allow us to reject the null hypothesis of neutral evolution for both types of sites.

*A few additional comments regarding different types of non-neutral mutations at nonsynonymous sites*

Let's assume some nonsynonymous mutations are neutral but others are selected.

(i) Very strongly deleterious mutations will not contribute to divergence and make a negligibly small contribution to polymorphism; one can think of them as not occurring at all (in this context) and it is like having a lower *neutral* mutation per site,  $\mu_{non,s} < \mu_{syn,s}$ . Because strongly deleterious alleles have little to no effect on both divergence and polymorphism, the ratio  $E(X_{non})/E(D_{non})$  is unaffected by them.

(ii) Weakly deleterious alleles make a negligible contribution to divergence but do add to polymorphism (such mutations segregate long enough to be observed as polymorphisms in within-species samples). Consequently,  $E(X_{non})/E(D_{non})$  will be larger than if all nonsynonymous mutations were neutral, i.e., negative selection will create an excess of polymorphism relative to divergence.

(iii) Beneficial mutations either are quickly lost by drift or sweep to fixation. Because they do not segregate for a long time, we are unlikely to “catch” them in this transitory state. We do not expect beneficial mutations to contribute to polymorphism but they will contribute to divergence. Consequently,  $E(X_{non})/E(D_{non})$  will be smaller than if all nonsynonymous mutations were neutral, i.e., positive selection will create an deficit of polymorphism relative to divergence.

Smith and Eyre-Walker (2002, *Nature*) made a simple extension of the McDonald-Kreitman (MK) test to estimate the fraction of fixed differences that were due to beneficial mutations. They assumed that a fraction  $f$  of nonsynonymous mutations were neutral and that the remainder were either strongly deleterious or beneficial (i.e., they assumed there were no weakly deleterious mutations). Then the expected amount of polymorphism is

$$E(X_{non}) = 4Nf\mu_{non}k_m$$

The expected amount of divergence is

$$E(D_{non}) = 2f\mu_{non}(G + 2N_A) + a$$

where the first term on the right reflects neutral divergence and  $a$  is the number of adaptive substitutions that have occurred since the two species split. Rearranging the equation above we have

$$a = E(D_{non}) - 2f\mu_{non}(G + 2N_A)$$

which can also be written as

$$a = E(D_{non}) - E(X_{non}) \frac{E(D_{syn})}{E(X_{syn})}$$

Dividing this expression by  $E(D_{non})$  gives

$$\alpha = 1 - \frac{E(X_{non})}{E(D_{non})} \frac{E(D_{syn})}{E(X_{syn})}$$

where  $\alpha$  is the fraction of nonsynonymous divergence that is due to beneficial mutations.

An important assumption here in this derivation is that there are no weakly deleterious mutations. How would such mutations affect our estimate of  $\alpha$  if we plugged in estimates of polymorphism and divergence into the equation above? (Hint: see (ii) above).

One simple way to reduce the problem of segregating deleterious mutations at nonsynonymous sites is to exclude sites with singletons (variants that appear only once among all the samples for a species). Deleterious variants will be over-represented among singletons so by ignoring such sites you remove many of these variants. To be fair, synonymous sites with singletons are also excluded. (If all mutations were neutral, both  $X_{non}$  and  $X_{syn}$  would be similarly reduced so the ratio in the  $\alpha$  equation above would be unaffected.)

More sophisticated methods have since been developed to allow weakly deleterious mutations that can contribute to polymorphism (and divergence) as well as demographic effects (e.g., changes in population size) that can affect cause  $E(X_{neutral})$  to be different from what it is expected under the “ideal” (constant size) case (Eyre-Walker & Keightley 2009 *Mol. Biol. Evol.*).

### Hudson-Kreitman-Aguadé Test

Consider two loci that differ in the amount of polymorphism they harbour. **If both loci are evolving neutrally**, this difference could be due to a difference in the mutation rate between the first and second locus (i.e.,  $\mu_1 \neq \mu_2$ ). Alternatively, there could be other explanations (e.g., one of the loci could be experiencing balancing selection or have recently experienced a selective sweep). To rule out the difference in mutation rate one



can compare the levels of polymorphism to the levels of divergence. For each locus, one must have multiple samples of the allele from the focal species (to measure polymorphism) as well as at least one allele from a related species (to measure divergence).

Analogous to our previous results, the ratio of polymorphism to divergence for locus 1 is

$$\frac{E(X_1)}{E(D_1)} = \frac{4N\mu_1 k_m}{2\mu_1(G + 2N_A)} = \frac{2Nk_m}{G + 2N_A}$$

Similarly, for locus 2,

$$\frac{E(X_2)}{E(D_2)} = \frac{2Nk_m}{G + 2N_A}$$

Again, under neutrality we expect  $\frac{E(X_1)}{E(D_1)} = \frac{E(X_2)}{E(D_2)}$ . A simple  $\chi^2$  test can be used to test for significant deviations.