

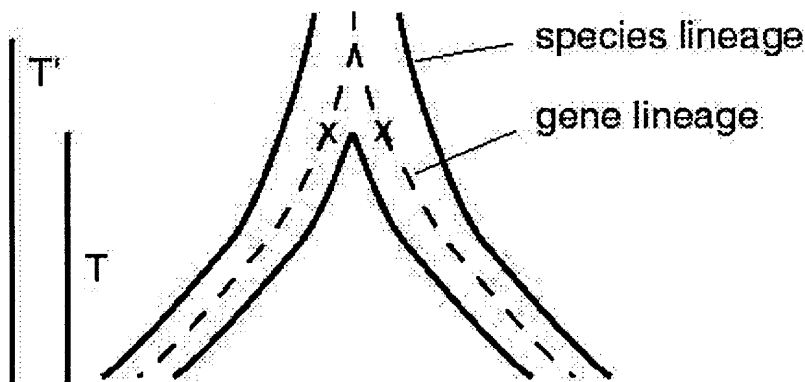
The HKA test (Hudson, Kreitman and Aguade. 1987. Genetics 116: 153-159) is a statistical test for determining whether increased polymorphism in a region is due to balancing selection. Because increased polymorphism may also be due to an elevated mutation rate in the region, it is necessary to rule out elevated mutation as an explanation. The HKA test does this by determining whether increased polymorphism within a species is accompanied by increased rates of silent substitution between species in the same region, since elevated mutation rates should also lead to greater rates of silent substitution.

The key result that forms the basis of this test is that under strict neutrality, both the proportion polymorphism in a region and the rate of substitution in that region are proportional to the mutation rate in the region. We have already seen that the former is the case, *i.e.* the proportion polymorphism is given by

$$P = 4 N \mu a_m ,$$

where N is population size, μ is mutation rate at neutral sites, and a_m is a constant associated with m copies of the gene being examined. In this handout, we derive from coalescence theory, the result that rate of substitution, as measured by fixed differences in the region between two lineages, is also proportional to mutation rate under strict neutrality.

4. We start with one allele of a gene from each of two sister species.
 - a. We know that these two alleles are descended from a common ancestor.
 - b. This common ancestral allele must have existed before speciation occurred.
 - c. Consequently, we can portray the events that led from this ancestral allele to the present copies by the following diagram:



- d. In this diagram, T represents the amount of time before present that the speciation event occurred, while T' represents the coalescence time for the two alleles.
- e. Now, suppose that we knew the sequences of these two alleles right at the time that speciation occurred (points marked with x on the diagram).

f. Assuming that this is a neutral sequence, we can use coalescence theory to determine what the expected proportion polymorphism is for these two alleles – note that this is the same as the proportion of sites that differ for the two alleles:

$$P_p = 4 N_a \mu a_m \quad ,$$

where $m = 2$ because there are two alleles sampled, and N_a is the effective population size of the ancestral species.

g. But $a_2 = \sum_{i=1}^1 \frac{1}{i} = 1$, so $P_p = 4 N_a \mu$. This is the expected proportion of neutral sites that differ between the two sequences just at the time of speciation. This means that the total number of differences in a sequence of γ base pairs is just

$$\text{Total No. of differences} = 4 N_a \mu \gamma \quad .$$

h. Now, after speciation, each sequence further accumulates mutations, and, since we are assuming all mutations are unique, each of these mutations increases the divergence between the two sequences.

i. The rate at which these mutations accumulate per site is just μ , so that the total number of such mutations in one species accumulated over time T at γ sites is simply $\mu T \gamma$. (Note that this is independent of the effective population size; see handout on molecular clock).

j. Consequently, the total number of fixed mutations accumulated in both species is simply $2\mu T \gamma$.

k. Putting this together with the previous result, the expected total number of differences accumulated between the two sequences is just:

$$\begin{aligned} \text{Exp. total No. differences} &= \text{No. differences accumulated prior to speciation} \\ &+ \text{No. differences accumulated subsequently} \\ &= 2\mu T \gamma + 4 N_a \mu \gamma \\ &= \mu \gamma (2T + 4 N_a) \end{aligned}$$

The expected proportion of sites showing fixed differences is thus

$$\begin{aligned} P_f = \text{Expected Proportion fixed differences} &= \frac{\text{Exp. No. fixed differences}}{\text{Number of sites}} = \frac{\mu \gamma (2T + 4 N_a)}{\gamma} \\ &= \mu (2T + 4 N_a) \end{aligned}$$

l. What this expression tells us is that for any sequence that behaves according to strict neutrality, the amount of sequence differentiation between species will be proportional to the rate of mutation in the region of that sequence.

m. Note that the parameters in the expression in parentheses will be the same for all sequences (genes) from the same two species, so that *if two genes are both not under selection, the amount by which they differ in between-species divergence will be proportional to the difference in mutation rates at the two genes.*

n. Consequently, for two neutral sequences, 1 and 2, we have

$$\frac{P_{f_1}}{P_{f_2}} = \frac{\mu_1(2T + 4 N_a)}{\mu_2(2T + 4 N_a)} = \frac{\mu_1}{\mu_2} \quad (1)$$

and similarly,

$$\frac{P_{p_1}}{P_{p_2}} = \frac{4N\mu_1}{4N\mu_2} = \frac{\mu_1}{\mu_2} \quad (2).$$

Putting (1) and (2) together yields

$$\frac{P_{f_1}}{P_{f_2}} = \frac{P_{p_1}}{P_{p_2}} \quad , \quad (3)$$

which should hold if the two sequences have evolved under strict neutrality. Moreover, recognizing that both sides of (3) can be multiplied by $\frac{\gamma}{\gamma}$ without changing the relationship, we have

$$\frac{\gamma P_{f_1}}{\gamma P_{f_2}} = \frac{\gamma P_{p_1}}{\gamma P_{p_2}} \quad , \text{ or}$$

$$\frac{\text{No. Fixed differences in sequence 1}}{\text{No. Fixed differences in sequence 2}} = \frac{\text{No. sites polymorphic in sequence 1}}{\text{No. sites polymorphic in sequence 2}} \quad (4)$$

o. This relationship thus serves as the null hypothesis that elevated levels of polymorphism at silent sites in sequence 1 are due to an elevated neutral mutation rate, μ_1 , compared to the mutation rate of sequence 2, μ_2 . The HKA test compares a sequence exhibiting elevated levels of polymorphism (sequence 1), with another sequence that is known or strongly suspected to be evolving neutrally (sequence 2). If the first sequence is evolving neutrally as well, and the excess polymorphism is due to a higher mutation rate, then (4) should be true. If (4) is not true, then the null hypothesis is rejected and it is concluded that the excess polymorphism in sequence 1 is due to balancing selection. A simple χ^2 test may be used to test the validity of (4).