# A new method to scan genomes for introgression in a secondary contact model-- Supplementary Material

## In this supplement, we derive a basic statistical description of the minimum coalescent time between samples taken from populations with no post-divergence gene flow. We then relate these coalescent time results to the minimum number of nucleotide differences between populations (, the numerator of *Gmin*) and to the behavior of the overall *Gmin* statistic under these (null-model) conditions.

## *Model*

We assume a single panmictic population of size  diploid individuals, split into two daughter populations (of size  and , respectively)  generations in the past with no post-divergence migration. In the present time, we have sampled  individuals from population 1 and  individuals from population 2. Let  be the number of generations before the present until the first coalescence between samples from different populations. Conditioning on  and  as the number of ancestral lineages remaining in each population at the time of divergence , we can write the total time until the first between-population coalescence as



 (S1)

Here,  is the time (looking backwards from ) until the first coalescent event in the pre-divergence population involving one of the  remaining lineages from (present-day) population 1 and one of the  remaining lineages from population 2, and is the probability that a sample of size  taken from population  has exactly  ancestors at time  generations before the time of sampling. We then have

 (S2a)

and

 (S2b)

as the mean and variance of the minimum between-population coalescence time under a null model of complete divergence. Expressions for  are well known [1]; we now turn to the derivation of the expected mean and variance of the time , which are less well understood.

*Coalescent analysis of *

In order to derive the mean and variance of , the minimum between-population coalescence time in the ancestral population, we condition on the number of coalescent events that have occurred in the ancestral population prior to the first coalescent event joining lineages from both subsamples. It is important to note here that we are only considering the genealogy in the ancestral (pre-divergence) merged population, as the  probabilities account for the genealogical history in the daughter populations post-divergence. Starting, then, from the time of the population split  and looking backwards, let be the time of the first between-population coalescent event given  prior ancestral coalescent events and  the probability of  ancestral coalescent events before the first between-population coalescent event. This yields

 (S3)

and

 (S4)

as expressions the mean and variance of the ancestral coalescent time.

From standard coalescent theory, we can readily write down , the expected time until a total of  coalescent events ( within-sample and one between-samples) have occurred in the ancestral merged population,

 (S5)

(where time is in units of generations), as well as the variance in that total time, given :

 (S6)

To obtain , begin with , the probability that the first coalescent is between samples, and , the probability that the first coalescent event is within a sample, immediately followed by the first between-sample event. These quantities can be written as

 (S7)

and



 (S8)

 was obtained by considering the two possible paths by which a single prior coalescence could have happened. A full general formula for  is obtained by extending this logic and summing over the possible paths by which prior coalescent events could have occurred,

 (S9)

The above expressions yield an expression for the expectation of the minimum between-population coalescence time from the time of divergence

 (S10)

and  can now also be obtained using equations S4, S6, and S9. The mean and variance of the overall expected minimum time to between-population coalescence under the null-model assumptions of population divergence and complete isolation at time  follow immediately from these results applied to equations S2a and S2b.

Relatively simple solutions for both mean and variance of  are also available for the special case in which only one ancestral lineage is present in at least one population at the time of divergence (**):

 (S11)

 (S12)

 (S13)

*Relation to* Gmin

The minimum between-population coalescence time is closely related to the numerator of the descriptive statistic *Gmin* ,. We cannot directly use the results above to develop a formal statistical description of *Gmin*, however, for two reasons. First, because *Gmin* is a ratio of two (related) random variables, a rigorous statistical treatment of *Gmin* requires derivation of quantities such as the covariance between numerator and denominator, currently unknown. Additionally, the relationship between minimum coalescence time and minimum number of mutations is complicated by the fact that samples with the minimum number of differences do not necessarily occur along the shortest genealogical branches of a tree. There is an extra level of sampling from the extremes of a stochastic process (mutation) that we have not accounted for in our treatment of the extreme value of the genealogical process. We expect this problem, leading to systematic biases in the expected mean and variance of the number of nucleotide differences under our null model, to be particularly acute for low mutation rates ().

Despite these difficulties, an application of a Poisson model of mutation to the minimum coalescent times derived above can provide some basic insights into the properties of the *Gmin* statistic. By assuming a Poisson model of mutation (with rate , where  for simplicity) along lineages, we can write down the following approximate expressions for the mean and variance of the numerator of *Gmin*,, in terms of the mean and variance of the minimum coalescent time and the other parameters:



 (S14)



 (S15)

We note that the resulting approximate E[] (obtained from substitution of our expressions for  and ) will be of order  while Var[] has both a and term. This is in accordance with the expectation and variance of the denominator of the *Gmin* , () as derived by Takahata [2], who gave the expectation as



The variance of the *Gmin* denominator as derived by Takahata [2] is more complicated, but has accessible approximations when  is very large (>>1),



or very small (<<1),



These results imply that the behavior of *G*min should be relatively insensitive to values of , consistent with our numerical simulations. The value of the *G*min ratio is, through , predominantly influenced by population factors affecting the number of lineages present in the samples at the time of divergence. We note that, for example, increasingly remote divergence times will result in small numbers of ancestral lineages, increasing the expected value of *G*min. In the limit of one lineage per sample,  is one, and the numerator and denominator of *G*min are equal in expectation. For the slightly less extreme case where there are two lineages remaining in one sample, , yielding a *G*min expectation close to one for greater than one. A bottleneck in one of the descendent populations will likely lead to a small number of lineages present at divergence, and a *G*min close to one is again expected. Thus, we posit that for a range of realistic population scenarios, the null expectation for *Gmin* should be extreme (approaching 1), giving us considerable potential to detect introgression events that constitute a departure from that null.

## Supplement References

1. Tavaré S. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol. 1984;26: 119-164.

2. Takahata N, Nei M. Gene genealogy and variance of interpopulational nucleotide differences. Genetics. 1985;110: 325-334.