Genomics and Bioinformatics

Matteo Fumagalli

March 2, 2020

This question is divided into five points (i-v), each one carrying equal weight.

We wish to study the evolution of spotted salamanders (*Ambystoma maculatum*) in Northern and Southern China. As such, we collected DNA samples from individuals from these two regions and obtained their genome. However, for practical reasons, we assess their genetic diversity only at three *loci*.

From Northern China, haplotypes for two diploid individuals over the three *loci* are
AAA
AAT
AGT
CAT
while for Southern China, two diploid individuals have haplotypes
AGT
AGT
AGT
AAT
.

We also have access to the putative ancestral haplotype over these three *loci*. The ancestral state of these *loci* is
AGT
.

(i)

Calculate the derived allele frequencies for the three *loci* for both populations.

(ii)

Calculate the unfolded site frequency spectrum for both populations.

(iii)

Estimate the population genetic parameter $\theta = 4N_e\mu$ using either Tajima's or Watterson's estimator for both populations.

(iv)

Assuming that both populations have been evolving under a scenario of constant population size and that their mutation rate is equal, discuss which population (Northern or Southern China) has a greater effective population size.

Explain how you reached your conclusions based on the results you obtained at both point (ii) and (iii).

(v)

The population genetic parameter $\theta = 4N_e\mu$ also represents the expected number of mutations separating two gene copies. Show that this is true when we measure time in $2N$ generations.

Hint: with a mutation rate of μ we expect ... mutations in r generations. If we measure time t in $2N$ generations, then this expectation becomes ... . We also know that the coalescence rate is 1 per $2N$ generations and there are two lineages separating the two gene copies. Therefore ...

Answers

(i)

For Northern China, the derived allele frequencies are 1/4, 3/4, 1/4. For Southern China, they are 0/4, 1/4, 0/4.

Here I test the students' understanding of concepts of haplotypes, allele frequencies, ancestral and derived state, using a working example. We used a very similar example in class.

(ii)

For Northern China, the unfolded SFS is $(0/3, 2/3, 1/3, 0/3)$ while for Southern is $(2/3, 1/3, 0/3, 0/3)$.

Here I test whether the students are able to calculate the frequency spectra and therefore move from a space of frequencies-per-site to proportions-of-sites as discussed and shown in class.

(iii)

With Tajima's estimator, this value would be $(1+2+2+1+1+2)/6 = 1.5$ for Northern and $(0+0+1+0+1+1)/6 = 0.5$ for Southern. With Watterson's, it is $3/(1/1 + 1/2 + 1/3) = 1.64$ for Northern and $1/(1/1 + 1/2 + 1/3) = 0.55$ for Southern.

Here I test whether students are able to provide an estimate of θ and understand the rational behing the chosen approach.

(iv)

Northern have a greater N_e given the formula of θ and that the mutation rate is the same. The sample size is too small to make inferences from the SFS but still Northern have more polymoprhic sites.

Here I test whether students understand the interpretation of θ.

(v)

With a mutation rate of μ we expect μr mutations in r generations. If we measure time t in $2N$ generations, then this expectation becomes $2N\mu t$. We also know that the coalescence rate is 1 per $2N$ generations and there are two lineages separating the two gene copies. Therefore $E[t] = 1$ and $2N\mu \times 2$ leading to $\theta = 4N_e\mu$.

Here I test for students who aim at a high Distinction whether they also have a formal understanding of the population genetic parameter θ.