Further comments on genetic diversity

M. Humberto Reyes-Valdés

May 30, 2019

CABANA workshop

Genomic analysis of crop diversity using R

Diversity is a general term referring to the existence of variation, which for the categorical case is basically the presence of several classes. There are multiple ways to measure diversity. In this note, I am concentrating on allelic diversity, as observed by the use of genetic markers.

I discuss further issues on genetic diversity based on DNA markers. My developments in this area have been based mainly in information theory, thus I will start from its core definition.

The mutual information between two variables X and M is defined as the average reduction in the uncertainty about X given knowledge of the value of M, in accordance with the following expression:

$$I(X;M) = H(X) - H(X|M) = H(M) - H(X|M) \tag{1}$$

In the heart of this definition is H, the uncertainty measured by the Shannon entropy. It turns out that the Shannon entropy is also a measure of available information. For the specific case of allelic diversity, the Shannon entropy is defined as follows:

$$H = -\sum_{i=1}^{k}(p_i)log_2(p_i), \tag{2}$$

where p_i is the allele frequency for the $i-th$ allele, and k the number of alleles in a given locus, with $(p_i)log_2(p_i) = 0$, for $p_i = 0$. The Shannon entropy can be used to measure diversity beyond allele frequencies, e.g. for genotypes and species. In ecology, the so-called Shannon diversity refers to this expression applied to species frequencies.

It has been argued (Jost et al., 2006) that the exponential entropy is a natural and better measure of diversity. It can be defined as follows:

$$D = 2^{(-\sum_{i=1}^{k}(p_i)log_2(p_i))} \tag{3}$$

Both, H and D tend to grow as the number of alleles increases, and as their frequencies reach uniformity.

A popular (and largely misused) measure of diversity is the so-called Polymorphism Information Content (PIC). Proposed by Botstein et al.(1980), it was developed for a kind of human pedigree, in which one parent is affected by a rare dominant disease and is heterozygous at the disease locus, whereas the other parent is unaffected by the disease. PIC aims to measure the information provided by this type of pedigree, extracted from a population in Hardy-Weinberg equilibrium, and is defined as follows:

$$PIC = 1 - \sum_{i=1}^{k} p_i^2 - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} 2(p_i p_j)^2 \tag{4}$$

Equation (4) cannot be considered a general measure of either diversity or information, because it is aimed to detect alleles linked to rare dominant diseases in the above described pedigree.

A general measure of diversity, often called PIC or heterocigosity, is the Gini-Simpson index, with the following expression:

$$GS = 1 - \sum_{i=1}^{k} p_i^2 \tag{5}$$

In this application, the Gini-Simpson index estimates the probability that two alleles randomly chosen are equal.

The Gini-Simpson index is bounded above by 1. However, the Shannon entropy does not have an upper bound, and grows indefinitely as the number of alleles increases.

R functions to perform the calculations of the so far described indices, can be found in Reyes-Valdés (2013).

A simple measure of allelic diversity is the so-called **allelic richness**. It is basically the number of alleles in a population.

The distance approach

When working with allele frequencies, average genetic distances among all pairs of individuals/lines in population sample, has also been used as an indicator of diversity.

The well-known euclidean distance is in general a good choice. For a given locus, it has the following expression:

$$D = \sqrt{\sum_{i=1}^{k} (p_i - q_i)^2}, \tag{6}$$

where p_i and q_i are allelic frequencies for each of two samples within a population. The formula for euclidean distance can be extended to m loci as follows:

$$D = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{k_i} (p_{ij} - q_{ij})^2}, \tag{7}$$

where p_{ij} and q_{ij} are allele frequencies for the $i-th$ locus and the $j-th$ allele, and k_i is the number of alleles in the $i-th$ locus.

A quite popular metrics of genetic distance is the Modified Roger's distance, defined as follows:

$$MR = \frac{1}{\sqrt{2m}} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{k_i} (p_{ij} - q_{ij})^2} \tag{8}$$

As one can see, the Modified Roger's distance is just a linear function of the euclidean distance.

Literature

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32:314–331.

Jost L, Baños T. 2006. Entropy and diversity. Oikos 113:373–375.

Reyes-Valdés MH. 2013. Informativeness of microsatellite markers. In: Microsatellites, Volume 1006 of Methods in Molecular Biology. S. Kantartzi (editor), p. 259–270. Springer.

$$\overbrace{\left(\hat{\mathrm{O}}_{\triangledown}\hat{\mathrm{O}}\right)}$$
$$\left(\left[\begin{array}{c}\mathrm{vvv}\\ \mathrm{vvv}\end{array}\right]\right)$$
$$\lambda\frown\lambda$$

Humberto Reyes