PolyPICker: Heterozygosity and Polymorphism Information Content

Genetic polymorphisms are tested for various reasons, but are most often used as markers to distinguish alleles or haplotypes carried in one individual with those of another, whether that other individual is a close relative, an unrelated member of the public, or an individual of a completely different ethnicity. The utility of a polymorphism in this regard (also referred to as its information content) is dependent on individuals being heterozygous for that polymorphism, and is, therefore, proportional to the probability that any randomly selected individual has a heterozygous genotype. Hence, one measure of a polymorphism's information content is its expected heterozygosity (or *H*): the proportion of a population that is expected to be heterozygous for the variant. If the frequency of each allele of a polymorphism is known, expected heterozygosity can be readily calculated using the principle of Hardy-Weinberg equilibrium as per the formula in Figure 1. This formula is essentially the sum of the probabilities of a homozygous genotype for each allele (e.g. $p^2 + q^2$ etc.) subtracted from 1.

$$H = 1 - \sum_{i=1}^{n} {p_i}^2$$

Figure 1: Formula for expected heterozygosity (*H*) based on the principle of Hardy-Weinberg equilibrium. *n* = number of alleles. p_i = frequency of the i^{th} allele.

The figure for *H* can range from 0 to 1 and is determined by two factors: the number of alleles and the distribution of frequency across those alleles. An invariant locus would, of course, have $H = 0$, as all individuals would be homozygous for the same allele. A rare biallelic variant (e.g. 1 heterozygote in 100,000 individuals) would have a *H* value of just above zero. The maximum value of *H* for a biallelic variant is 0.5 (i.e. $1 – [0.5^2 + 0.5^2]$). The value of *H* is maximal for markers that have a large number of alleles (*n*) where the frequency assigned to each allele is distributed as close to equally (i.e. 1 / *n*) as possible.

When evaluating genetic markers in family linkage studies, a meiosis is only considered fully informative if the parental origin of each allele detected in offspring is clear and unambiguous. For this to be the case when an offspring is heterozygous, each parent must have a different genotype. For example, a parent-offspring trio in which all three are heterozygous for the same two alleles is not informative, as which allele was inherited from which parent cannot be determined unambiguously.

The probability of a parent-offspring trio all having the same heterozygous genotype (hereafter referred to as *TH*) is calculated by multiplying the probability of each possible heterozygous genotype (e.g. 2*pq*, as determined by Hardy-Weinberg equilibrium principle) by itself (i.e. $[2pq]^2$; this is the probability of the two parents having the same heterozygous genotype) and then by 0.5 (i.e. the probability of the parents having an offspring that also has the same genotype). This equates to $[2pq]^2 \times 0.5 = 2 \times p^2 \times q^2$. The values obtained for each possible heterozygous genotype for the marker in question are then summed to give the overall *TH* probability. The formula for *TH* is summarised in Figure 2.

$$TH = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2{p_i}^2 {p_j}^2$$

Figure 2: Formula for the probability that all three members of a randomly selected parent-offspring trio will have the same heterozygous genotype (*TH*). *n* = number of alleles. p_i = frequency of the i^{th} allele. p_j = frequency of the j^{th} allele.

Given the possibility of a heterozygote being uninformative due to having parents with the same heterozygous genotype, the *H* value is not an ideal measure of a marker's information content for linkage studies. However, this shortcoming can be resolved by simply subtracting *TH* from *H* to give a figure known as polymorphism information content (*PIC*). The *PIC* value can be interpreted as the probability of a randomly selected individual having a heterozygous genotype that is not the same as both of his / her parents. This adjusted value of *H* was proposed by Botstein *et al*. (1980) as a more useful means to select informative markers specifically for linkage studies. In their article, the information content of markers was categorised as low, medium, or high for those with *PIC* values of <0.25, 0.25 – 0.5, or >0.5 respectively. The formula for *PIC* can be summarised as *PIC* = *H* – *TH* and is displayed in full in Figure 3.

$$PIC = 1 - \sum_{i=1}^{n} {p_i}^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2{p_i}^2 {p_j}^2$$

Figure 3: Formula for polymorphism information content (*PIC*) as described by Botstein *et al*. (1980). *n* = number of alleles. p_i = frequency of the i^{th} allele. p_j = frequency of the j^{th} allele.

The calculation of *H* and *PIC* can be quite an arduous task if done by hand, especially for polymorphisms with large numbers of alleles. PolyPICker is an online program (https://www.genecalculators.net/pq-chwe-polypicker.html) that has been designed to quickly calculate *H* and *PIC* using the formulae shown in Figures 1 and 3. To use this program, simply select the number of variants (1 – 5), the number of alleles for each variant (2 – 20), and then input frequency values for each allele in the yellow boxes (note that the total allele frequency for each variant must add up to 1). The *H* and *PIC* values, expressed as a percentage, will then be displayed.

References:

Botstein *et al*. (1980). *Am J Hum Genet*, 32(3): 314-331. PMID: 6247908