Hardy-Weinberg equilibrium and linkage disequilibrium

Population refers to a group of organisms of the same species living in a sufficiently restricted geographical area so that any member can potentially mate with any other member of the opposite sex. Population can be subdivided into local inbreeding units by environment factors or social behaviours. Individuals are likely to choose their mates from the same subdivided population. Therefore, such subdivided population is refered as local inbreeding units or local populations. It is at the level of local population that adaptive evolution takes place through systematic changes in allele frequency. Therefore, in population genetics, local population is refered as population, and are sometimes refered as Mendelian population or subpopulation.

1 Random mating and nonoverlapping generations

In sexual organisms, genotypes are broken into gamete and assembled in fertilization. The proportion of a specific genotype in a population is genotype frequency. The formation of a genotype in new generation is determined by the opportunity for gametes to come together, and the opportunity for gamete to come together is determined by mating that take place. Therefore, genotypes of mating pairs determines genotypes of offsprings. In population genetics, random mating is a widely used idealization in which mating pairs have the same frequencies as if they were formed by random collisions between genotypes. In real world, random mating is often complicated by trait preference and population structure.

Nonoverlapping generation model is wodely used in population genetics. It assumes that the cycle of birth, maturation and death includes the death of all individuals in each generation before the members of next generation mature. It literally applies to short-living organisms like some insects and annual plants.

2 Hardy-Weinberg equilibrium (HWE)

Hardy-Weinberg equilibrium is based on following assumptions: (1) diploid organisms; (2) sexual reporduction; (3) nonoverlapping generations; (4) the gene under consideration has two alleles; (5) identical allele frequencies in males and females; (6) random mating; (7) large (infinite) population size; (8) no im-

migration; (9) no mutation; (10) no natural selection. Denote by A and a two alleles considered, the population is of HWE if the genotype frequencies of AA, Aa and aa is p^2, $2pq$ and q^2, where $p+q=1, 0<p<1$. This relationship between allele frequencies and genotype frequencies is summarized as

$$(p+q)^2 = p^2 + 2pq + q^2 \tag{1}$$

It represents such a fact: under HWE, genotype is resulted from random combination of two alleles, or random combination of gametes.

HWE serves as a null model in which no evolutionary force driving change of allele frequencies are imposed. It affords a baseline for comparison with more realistic models in which evolutionary forces are imposed. Also, HWE separates life history into two intervals: gametes combing into zygotes and zygotes producing gametes. In construction of more realistic models, one can often introduce complications into zygote-to-gamete part. Obviously, allele/genotype frequencies of a population in HWE are not changed after one generation of random mating. Also, any given genotype frequencies goes into HWE in one generation of random mating since random mating is equalivant to random union of gametes. Even with overlapping generations, HWE is reached after a relatively small number of generations.

3 Testing for HWE

HWE can be tested by goodness-of-fit test. χ^2 value is given by

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \sim \chi^2(df) \tag{2}$$

where *observed* is the oberved genotype frequencies, *expected* is the expected genotype frequencies given by HWE and gene frequencies from *observed*. χ^2 follows chi-square distribution with degree of freedom df. In the case of one gene with two alleles, two allele frequencies is estimated from sample with the relationship that the sum equals 1, so the degree of freedom is $2-1=1$.

Generally, *expected* values in Eq. 2 should be above 5. When sample size is so small that it violate the convention but all *expected* values are not too close to 0, Eq. 2 can be modified as

$$\chi^2 = \sum \frac{(|observed - expected| - 0.5)^2}{expected} \sim \chi^2(df) \tag{3}$$

When sample size is small, exact test for HWE can be used. Consider a sample of size n with n_{11} individuals of AA, n_{12} of Aa and n_{22} of aa, the number of allele A and a in the sample are $N_1 = 2n_{11}+n_{12}$

and $N_2 = n_{12} + 2n_{22}$. The probability of n_{12}, conditional on N_1 and N_2 is given by

$$\Pr\{n_{12}|N_1, N_2\} = \frac{n!/(n_{11}!n_{12}!n_{22}!)}{(2n)!/(N_1!N_2!)} 2^{n_{12}} \tag{4}$$

P value of observed n_{12} is given by cumulative probability of all possible n_12 values with probability below the observed n_{12} value. Generally, reject HWE when $P > 0.05$.

Permutation test can also be used. A random permutation of $N_1 + N_2$ alleles with form $(A, a, \ldots, a, A)$ can be generated. From a permutation, a configuration of genotypes (n_{11}, n_{12}, n_{22}) can be generated by combining $(2n-1)$th allele and $2n$th allele. Then, a P value can be calculated by goodness-of-fit test comparing the configuration of genotypes with HWE. Repeat this process multiple times (¿1000), the significant level of permutation test can be given by proportion of P values that are below 0.05.

Dominant alleles also complicate testing for HWE since it obscures one-to-one relationship between phenotypes and genotypes. Let A be dominant regards to a. Given a sample of n individuals with genotype frequency R for aa, gene frequency of a under HWE is estimated by

$$\hat{q} = \sqrt{R} \tag{5}$$

Standard error of $\hat{q}$ is given by

$$SE(\hat{q}) = \sqrt{\frac{1-R}{4n}} \tag{6}$$

Under HWE, the genotype frequency of heterozygotes Aa is $2pq$, which maximizes when $p = q = 0.5$. The ratio of genotype frequencies of Aa and aa is

$$\frac{2pq}{q^2} = 2(\frac{1}{q} - 1) \tag{7}$$

which increases as q goes towards 0, leading to the excess of heterozygotes over recessive homozygotes aa.

4 Extension of HWE

HWE can be extented to multiple alleles. Consider a gene with multiple alleles $A_1, A_2, \ldots, A_n$, and let the allele frequency of A_i be p_i, the HWE genotype frequency of A_iA_i is p_i^2, and frequency of A_iA_j $(i \neq i)$ is $2p_ip_j$. Such relationship is summarized as

$$(p_1 + p_2 + \cdots + p_n)^2 = p1^2 + p_2^2 + \cdots + p_n^2 + \sum_{i,j} 2p_i 2p_j \tag{8}$$

Genes on sexual chromosomes can be modeled by HWE. Consider a X-linked gene with two allele X^A and X^a with allele frequencies p and q, and the allele frequencies in male and female subpopulations are identical. The genotype frequencies of females are p^2 for X^AX^A, $2pq$ for X^AX^a, and q^2 for X^aX^a. Regards to males, genotype frequencies are p for X^AY and q for X^aY. As a result, if A is dominant to a, the recessive genotype is more common in males than females. Obviously, for X-linked genes, allele/genotype frequencies of a population in HWE are not changed after one generation of random mating. Also, any given genotype frequencies goes into HWE in one generation.

5 Linkage equilibrium

Consider one gene with allele A and a, and another gene with allele B and b, and both genes are under HWE. If allele A/a randomly associates with allele B/b, *i.e.* gamete frequency equals multiplication of corresponding allele frequencies, the two genes are of linkage equilibrium, otherwise of linkage equilibrium.

Consider heterozygote genotype AB/ab, it produces four kinds of gametes (haplotype): AB, ab, Ab and aB. AB and ab are nonrecombinant gametes while Ab and aB are combinant gametes. AB and ab have equal frequencies, and Ab and aB have equal frequencies. The frequency of recombination r refers to the proportion of recombinant gametes produced by a double heterozygote. r is dependent on the distance of two genes. If two genes are on different chromosome, $r = 0.5$. If they are on the same chromosome, the closer they are, the lower the frequency of recombination is.

Consider a population with gene A/a and B/b, and let all assumptions for HWE model holds except the assumption of no recombinantion. Denote by r the frequency of recombination between two genes, and p_A, p_a, p_B, p_b allele frequencies. Allele frequencies do not change along generations due to random mating and infinite population size. The frequency of one haplotype in offspring generation ($(n+1)$th) is given by following relationship with the corresponding haplotype frequency in parent generation (nth),

$$P_{n+1}^{ij} = (1-r)P_n^{ij} + rp_ip_j \tag{9}$$

where $i = A, a$ and $j = B, b$. Therefore,

$$P_{n+1}^{ij} - p_ip_j = (1-r)(P_n^{ij} - p_ip_j) \tag{10}$$

Let $D_n^{ij} = P_n^{ij} - p_i p_j$, and

$$\begin{aligned}\frac{D_{n+1}^{ij}}{D_n^{ij}} &= \frac{P_{n+1}^{ij} - p_i p_j}{P_n^{ij} - p_i p_j} \\ &= \frac{(1-r)P_n^{ij} + r p_i p_j - p_i p_j}{P_n^{ij} - p_i p_j} \\ &= 1 - r\end{aligned} \tag{11}$$

Therefore, let n begin from 0,

$$D_n^{ij} = (1-r)^n D_0^{ij} \to 0, n \to \infty \tag{12}$$

When n is large, $D_n^{ij} \approx 0$ and $P_n^{ij} \approx p_i p_j$, linkage equilibrium is reached. The rate towards linkage equilibrium is dependent on r.

Note the following facts:

$$\begin{aligned}P_n^{AB} + P_n^{Ab} &= p_A \\ P_n^{aB} + P_n^{ab} &= p_a \\ P_n^{AB} + P_n^{aB} &= p_B \\ P_n^{Ab} + P_n^{ab} &= p_b \\ p_A + p_a &= 1 \\ p_B + p_b &= 1\end{aligned} \tag{13}$$

Therefore, for all four haplotypes,

$$D = D_n^{AB} = -D_n^{Ab} = D_n^{ab} = -D_n^{aB} = P_n^{AB} P_n^{ab} - P_n^{Ab} P_n^{aB} \tag{14}$$

Therefore,

$$\begin{aligned}P_n^{AB} &= D_n^{AB} + p_A p_B = D + p_A p_B \\ P_n^{Ab} &= D_n^{Ab} + p_A p_b = -D + p_A p_b \\ P_n^{ab} &= D_n^{ab} + p_a p_b = D + p_a p_b \\ P_n^{aB} &= D_n^{aB} + p_a p_B = -D + p_a p_B\end{aligned} \tag{15}$$

Thus, D provides a measure for linkage equilibrium and is known as linkage disequilibrium parameter or coefficient of linkage disequilibrium.

Other two widely-used measure of linkage equilibrium can be derived from D. Since haplotype frequency cannot be negative, the maximum and minimum of D is given by

$$\begin{aligned}D_{\min} &= \max\{-p_A p_B, -p_a p_b\} \\ D_{\max} &= \min\{p_A p_b, p_a p_B\}\end{aligned} \tag{16}$$

D' is defined as

$$D' = \frac{D}{D_{\min}}, D \leq 0$$
$$D' = \frac{D}{D_{\max}}, D > 0 \tag{17}$$

r^2 is defined as

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} \tag{18}$$