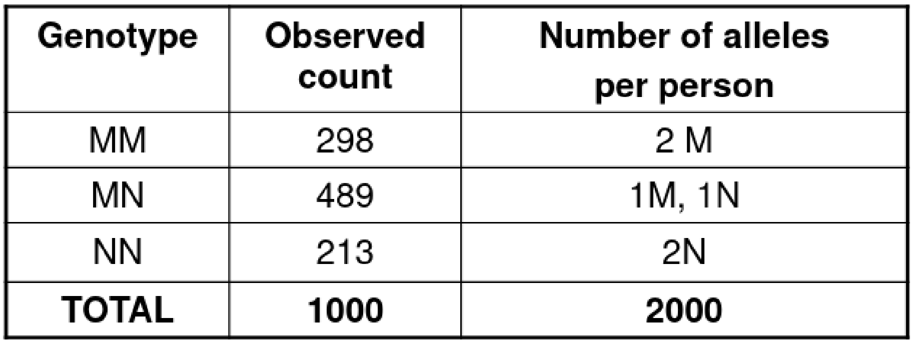
Population Genetics

Genotype and Allele Frequency Estimation is the first step in studying a polymorphism. Used for family data and independent individuals in a population. We can use a subset of individuals who are independent and count alleles, or use the maximum likelihood methods to take all genotypes into account for pedigree data.

Consider the following example of allele frequency estimation:



Genotype	Observed count	Number of alleles per person
MM	298	2 M
MN	489	1M, 1N
NN	213	2N
TOTAL	**1000**	**2000**

We can take the frequency of each allele by the observed proportions:

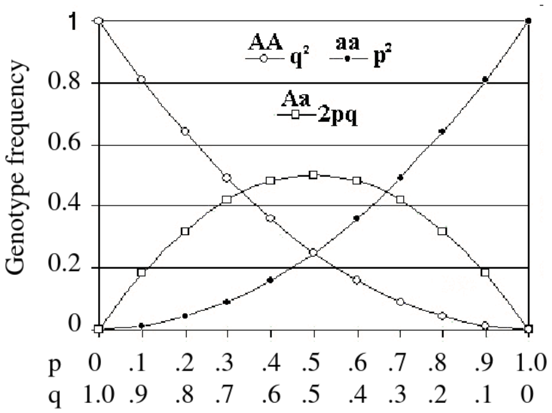
p_M = (2*298 + 489)/(2*1000)
p_N = 1 - pM

Hardy-Weinberg Law

Describes how we expect allele frequencies and phenotype frequencies to be related in a population.

1. For a large, random-mating population, in the absence of forces that change allele frequencies, *the allele and genotype frequencies remain constant from one generation to the next*
2. After one generation of random mating, for an autosomal locus with alleles 1 and 2 (frequencies p and q = 1 - p), the relative frequencies of the genotypes 11, 12, and 22 are:

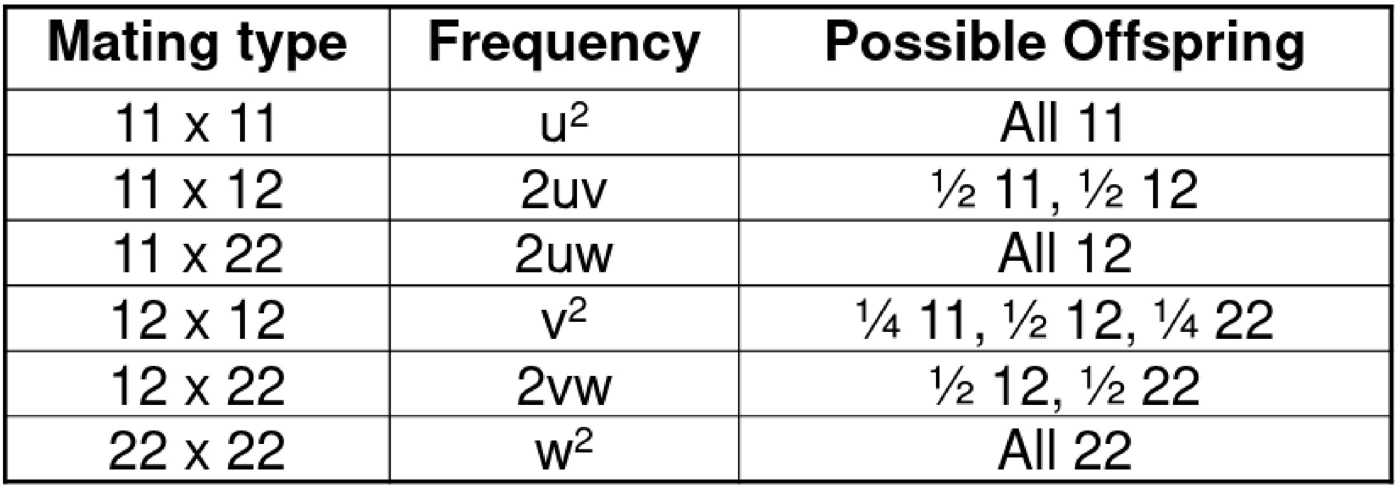
   p^2, 2pqm q^2



Assumptions

- Random mating with respect to genotype
    - No assortative mating
    - No population structure
- No selection, mutation, or migration
- Discrete generations
- Infinite population size
- Autosomal locus

Given two alleles with 1 and 2, there are 6 possible parent mating types:



Mating type	Frequency	Possible Offspring
11 x 11	u^2	All 11
11 x 12	2uv	½ 11, ½ 12
11 x 22	2uw	All 12
12 x 12	v^2	¼ 11, ½ 12, ¼ 22
12 x 22	2vw	½ 12, ½ 22
22 x 22	w^2	All 22

So the frequency of allele 1 in the offspring would be:
P(11) + ½P(12) = (u+v/2)2 + (u+v/2)(v/2+w) = (u+v/2)(u+v+w) = **u+v/2**

Similarly, the frequency of allele 2 is: **w + v/2**

Forces that change allele or genotype frequency (invalidate HW law)

- Mutation
- Migration
- Selection
    - Deleterious mutations tend to be rare if there is selection against them
        - Exception: Heterozygote advantage for a recessive deleterious

- Drift - small populations
- Non-random mating

Testing for HWE

Though several assumptions of the HW law are not met in any population, genotypes in a population usually conform reasonable well to expectations, due to the various forces cancelling each other out.

H0: The genotype frequencies math the HW expectations (p^2, 2pqm q^2)

1. Estimate allele frequency (p_hat)
2. Determine the *expected* genotype frequencies from the estimated allele frequency, assuming null is true

   $$(\hat{f}_{AA}, \hat{f}_{Aa}, \hat{f}_{aa}) = (\hat{p}^2, 2\hat{p}\hat{q}, \hat{q}^2)$$
3. Compute the *expeceted* genotype counts

   $$E_i = N \times (\hat{f}_{AA}, \hat{f}_{Aa}, \hat{f}_{aa})$$
4. Compare *observed* genotype counted to *expected*

   $$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Compare X^2 to a chi-squared distribution with 1 degree of freedom. This is usually the number of categories minus 1, but we lose an additional degree of freedom since we estimate allelic frequencies from the data (3-1-1).

Conclusion: The observed genotype frequencies are [not] significantly different from the expectations of the HW equilibrium.

When we reject the HWE, we usually don't know why other than the assumptions being violated.

Exact HWE Test

There are $(2N)!/n_A!n_B!$ possible arrangements for the alleles in the sample. Under HWE the probability of observing exactly n_{AB} heterozygotes in N individuals with nA alleles is:

$$P(N_{\mathrm{AB}} = n_{\mathrm{AB}} | N, n_{\mathrm{A}}) = \frac{2^{n_{\mathrm{AB}}} N!}{n_{\mathrm{AA}}! n_{\mathrm{AB}}! n_{\mathrm{BB}}!} \times \frac{n_{\mathrm{A}}! n_{\mathrm{B}}!}{(2N)!}$$

Under many conditions, samples of affected individuals will not be in HWE for alleles associated with disease **BUT** controls should be close to HWE, as should population-based (unascertained) samples. Note also that genotypes among related individuals may not be in HWE since the individuals are not independent.

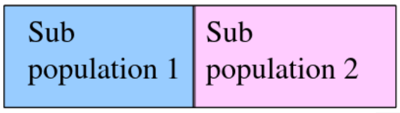
If the observed genotypes in unaffected controls (whole random sample) depart from HWE, this may indicate:

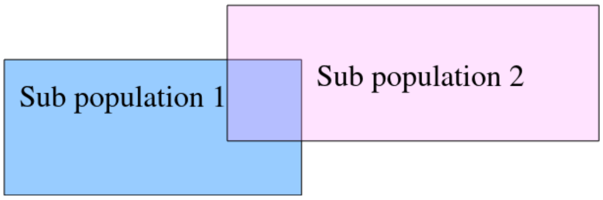
- Bad Assay - non-random genotyping error or non-random missing data
- Population structure (non-random mating)

Often if the HWE test p-value is << .01 in controls, the marker is not used in association analyses.

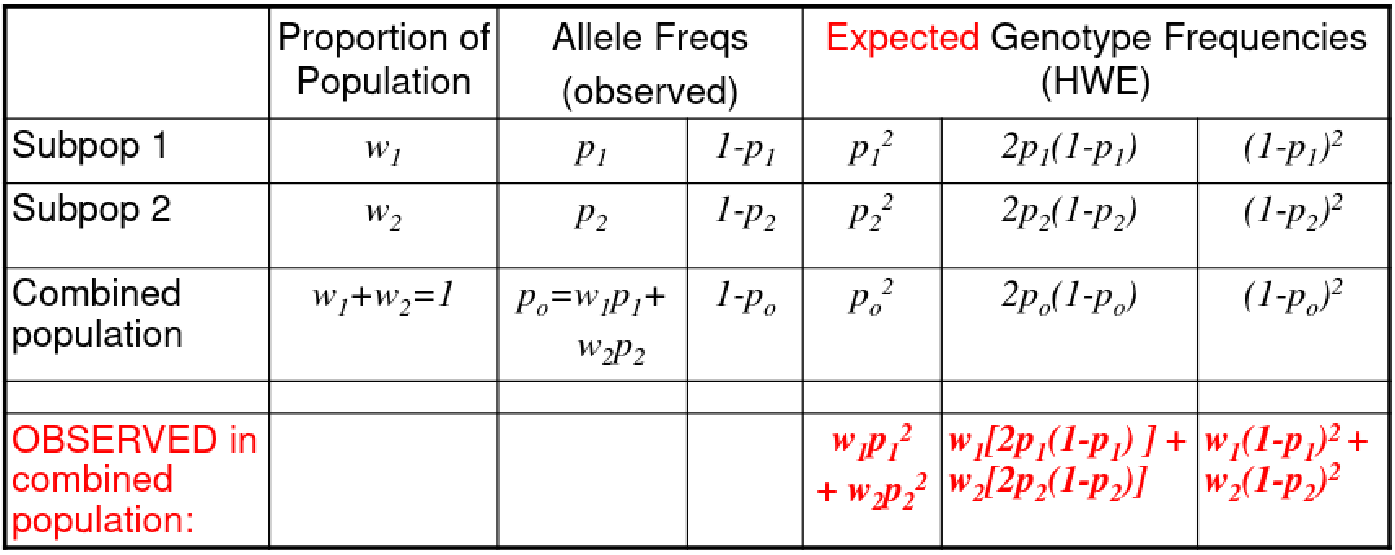
Population Structure

"Random Mating" occurs within a population, but not within overall population.





We can observed the combined population to test for structure in sub-populations:



	Proportion of Population	Allele Freqs (observed)		Expected Genotype Frequencies (HWE)		
Subpop 1	w_1	p_1	$1-p_1$	p_1^2	$2p_1(1-p_1)$	$(1-p_1)^2$
Subpop 2	w_2	p_2	$1-p_2$	p_2^2	$2p_2(1-p_2)$	$(1-p_2)^2$
Combined population	$w_1+w_2=1$	$p_o=w_1p_1+w_2p_2$	$1-p_o$	p_o^2	$2p_o(1-p_o)$	$(1-p_o)^2$
OBSERVED in combined population:				$w_1p_1^2 + w_2p_2^2$	$w_1[2p_1(1-p_1)] + w_2[2p_2(1-p_2)]$	$w_1(1-p_1)^2 + w_2(1-p_2)^2$

The Heterozygote Deficit = 1 - (observed het freq)/(HWE expected het freq)

$$F = 1 - \left[2w_1p_1(1-p_1) + 2w_2p_2(1-p_2)\right] \Big/ \left[2p_o(1-p_o)\right]$$

F is the proportional decrease in heterozygotes observed under what would be expected under the HWE.

Linkage Disequilibrium (LD)

Recall a haplotype is a set of markers on the same chromosome that are always inherited together. The haplotype consists of two pieces of information: Genotypes and which alleles are inherited together..

Suppose we have two markers 1, with alleles A and a and freq p_A, $p_{a,}$ and 2, with alleles B and b and freq p_B, p_b
We have 4 possible haplotypes: AB, Ab, aB, ab
If p_{AB} is the probability A and B are on the same chromosome, then we can say $p_{AB} = p_B * p_b$ **if markers are independent**
Let P_t(AB) be the frequency of the haplotype AB after t generations of random mating, and theta is the recombination fraction

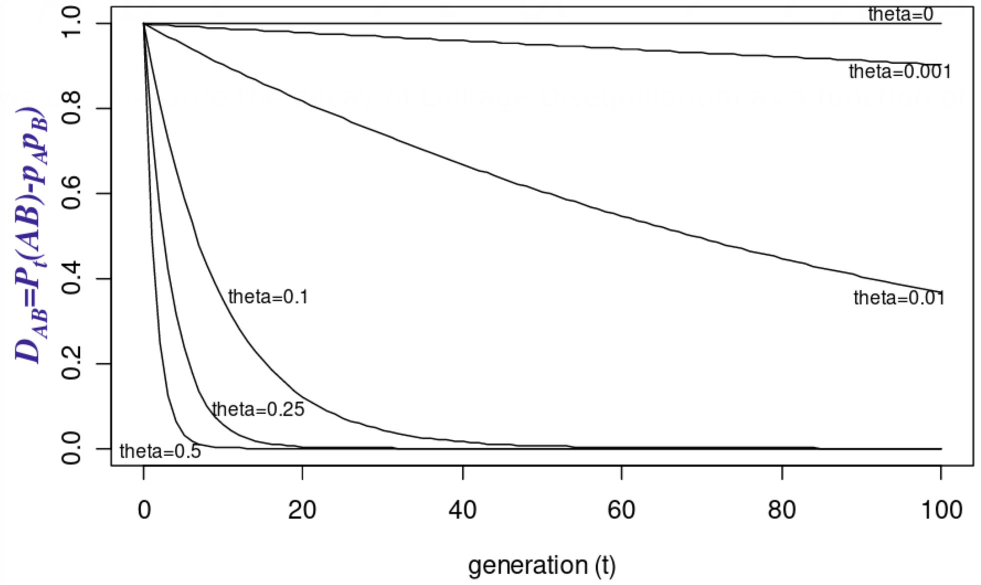
$$P_t(AB)=(1-\theta)\ P_{t-1}(AB) + \theta p_A p_B$$

$$P_t(AB)-p_Ap_B= (1-\theta)\ [P_{t-1}(AB) - p_Ap_B]$$

by induction,

$$P_t(AB)-p_Ap_B=(1-\theta)^t\ [P_0(AB) - p_Ap_B]$$

The idea being that



generations passed:

You can also measure this in a pairwise fashion:

$$D_{AB}=p_{AB}-p_Ap_B \qquad or\ p_{AB}=p_Ap_B+D_{AB}$$

Similarly,

- □ $p_{Ab}=p_Ap_b-D_{AB}$
- □ $p_{aB}=p_ap_B-D_{AB}$
- □ $p_{ab}=p_ap_b+D_{AB}$

There are problems with the pairwise measure however:

- The sign of D_{AB} is arbitrary
- Range depends on allele frequencies
- Can't easily compare the different pairs of markers

Min and Max of D_{AB}:
If D>0, $D \leq \min(p_Ap_b,p_ap_B)$
If D<0, $|D| \leq \min(p_Ap_B,p_ap_b)$

D is usaully scaled so that its range is (0, 1) or (-1, 1):

$$D'=\begin{cases} \dfrac{|D|}{\min(p_Ap_b,p_ap_B)} & \text{if } D>0 \\[2ex] \dfrac{|D|}{\min(p_Ap_B,p_ap_b)} & \text{if } D<0 \end{cases}$$

$$r^2=\frac{D^2}{p_Ap_ap_Bp_b}$$

Range of D' and r^2 is 0-1

D' = 1 -> No evidence for recombination beteween markers
D' = 1 -> Fewer than 4 haplotypes are observed between two biallelic variants
If allele frequencies are similar D' near 1 -> the markers are good surrogates

D' estimates can be inflated with small sample sizes
D' estimates can be inflated when one allele is rare

Revision #3
Created 3 October 2022 22:01:13 by Elkip
Updated 4 October 2022 00:51:24 by Elkip