

**How to test if HW assumptions hold.**

- Consider a gene with two types of alleles: A and a
  - $P_A + P_a = 1$  <--- allele frequencies. Based on HW, the frequency of genotypes are AA:  $P_A^2$ ; Aa:  $2P_AP_a$ ; aa= $P_a^2$ . How can we test if HW holds? From n individuals (2n chromosomes), say  $n_A$  have A allele and  $n_a$  have a allele.
  - Step 1: Find allele frequencies
    - $P_A = n_A/2n$ ;  $P_a = n_a/2n$
  - Step 2: Find expected numbers of individuals for each genotype
    - Expected frequency of AA genotype =  $P_A^2$
    - Expected number of individuals with AA =  $P_A^2 * n$ ,
    - Use similar calculations to find expected number with genotypes Aa and aa.
- What if observed values are different than expected values? Test to quantify if HW holds.
  - 1<sup>st</sup> option: Chi-Squared test.
    - Compute chi-squared value ( $\sum(o-e)^2/e$ ), to get p-value. p-value is the probability that you see more skewed distribution than observed if the HW assumption holds.
  - 2<sup>nd</sup> option: "Exact Test" more accurate than chi-squared.
    - Example: Suppose there are 8 people (16 chromosomes) with  $n_A=8$  and  $n_a=8$ . There are five possible distributions of genotypes ( $n_{AA}, n_{Aa}, n_{aa}$ ):
      - $P1=(0,8,0)$ ;  $P2=(1,6,1)$ ;  $P3=(2,4,2)$ ;  $P4=(3,2,3)$ ;  $P5=(4,0,4)$
    - For each of the  $P_i$ 's we want to compute  $\text{Prob}(n_{Aa} | n_A, n_a) = P(n_{Aa}, n_A, n_a) / P(n_A, n_a)$ 
      - = # cases with  $(n_{Aa}, n_A, n_a)$  / # cases with  $(n_A, n_a)$
      - To compute denominator, we are dealing with 16 chromosomes. Count cases with  $n_A$  and  $n_a$ . Sample case:  $n_A$  A's in a row followed by  $n_a$  a's in a row. There are  $(2n)! / (n_A)! (n_a)!$  such cases
      - To compute numerator, we are dealing with 8 individuals. Use the same principal as above, but keep in mind that  $Aa! = aA$ , so multiply the number by  $2^{n(Aa)}$ . This comes to  $n! / n_{AA}! n_{Aa}! n_{aa}! * 2^{n(Aa)}$
    - Do above calculation for each of the  $P_i$ 's. Compute p-values for each  $P_i$  by summing  $P_i$ 's with more or equally skewed distributions. Set some threshold p-val like 5% or 1%.

- **Properties of pairs of alleles; haplotype frequencies and linkage (dis)equilibrium.**
- First, consider the history of two alleles (Lecture 10, slides 4-6)
- Linkage disequilibrium: Some large part of a chromosome are conserved.
- Consider loci A, B. A has A or a types. B has B or b types.  $P_{AB}=P_A P_B$  if loci are independent. This assumption is linkage equilibrium, and occurs when there is frequent crossing over between loci, or if they are on different chromosomes.
- Why is this important? Cheaper to genotype: targeted sequencing. Set markers to infer genotype. (Lecture 10, slide 9). Some combinations of alleles happen a lot more frequently than others. There are recombination “hot spots.” Most loci are not hot spots, so they tend to stick together.
- A way of measuring linkage disequilibrium.
  - Haplotype frequencies: a certain assignment of types of chromosomes
  - Consider sites A and B and their haplotype frequencies (Lecture 10, slide 12)
  - Assume in LE, ie that  $P_{AB}=P_A P_B$ . (and other equations on Lecture 10, slide 13)
  - If in disequilibrium, then the equalities won't hold
  - Let  $D_{AB}$  = difference between the observed and expected values.
  - Use this to find equations on slide #15