Lecture 2. Genetic variation. Allele and genotype frequencies. Random mating and Hardy-Weinberg equilibrium.

1.1 Genetic variation

Two measures of genetic variation

Polymorphism = proportion of polymorphic genes
  with most common allele frequency $p \leq 0.95$
Heterozygosity = proportion of heterozygous genes
  in an average individual

Ex 1: numerical example

Next table gives an example of a sample of
  four individuals with Pm $= 0.3$, and $\bar{H} = 0.1$
Assignment
  1) explain the meaning of the ratio $\frac{\bar{H}}{Pm}$
  2) using the same format suggest two other samples
  with Pm $= 0.1$, $\bar{H} = 0.1$ and Pm $= 1.0$, $\bar{H} = 0$

Genes	1*	2	3*	4	5	6	7	8	9	10*	H_{ind}
Ind. 1	+	+	+	+	+	+	+	+	+	+	
	−	+	+	+	+	+	+	+	+	+	0.1
Ind. 2	+	+	−	+	+	+	+	+	+	+	
	+	+	+	+	+	+	+	+	+	+	0.1
Ind. 3	−	+	+	+	+	+	+	+	+	+	
	+	+	+	+	+	+	+	+	+	−	0.2
Ind. 4	+	+	−	+	+	+	+	+	+	+	
	+	+	−	+	+	+	+	+	+	+	0
$\bar{H}$	0.5	0	0.25	0	0	0	0	0	0	0.25	$H = 0.1$

Ex 2: allozyme polymorphisms

Fig 2.9, p. 55: 14 to 71 genes (mostly ≈ 20) in 243 species
  overall $x \pm s$: Pm $= 0.26 \pm 0.15$, $\bar{H} = 0.07 \pm 0.05$
Drosophila species - most polymorphic group
  mammals - least variable
  cheetah almost monomorphic

$$\bar{x} := \frac{x_1 + \ldots + x_n}{n}, \quad s^2 := \frac{1}{n-1}\Sigma_{i=1}^n (x_i - x)^2$$

Ex 3: nuclear DNA polymorphisms

Alcohol dehydrogenase (Adh) in $D.melanogaster$
  Fig 2.10, p. 58: 93 out of 113 alleles
Only two 2 allozymes due to a single
  nonsynonimous mutation at amino acid number 193
      slow allozyme $Adh\text{-}S$: AAG $=$ Lysine,
      fast allozyme $Adh\text{-}F$: ACG $=$ Threonine
  fast allele is more active and expressed

1.2 Allele and genotype frequencies

one locus two allele model of a diploid population

Diploid population size N

genotype counts $N_{AA} + N_{Aa} + N_{aa} = N$

Haploid population size $2N$

allele counts $(2N_{AA} + N_{Aa}) + (2N_{aa} + N_{Aa}) = 2N$

Genotype frequencies

$D = \frac{N_{AA}}{N}$, $H = \frac{N_{Aa}}{N}$, $R = \frac{N_{aa}}{N}$

$D + H + R = 1$

Allele frequencies

$p = \frac{2N_{AA}+N_{Aa}}{2N} = D + \frac{H}{2}$, $q = \frac{2N_{aa}+N_{Aa}}{2N} = R + \frac{H}{2}$

$p + q = 1$

$$\boxed{\begin{array}{c} D = p^2 + pqF, \ R = q^2 + pqF, \ H = 2pq(1 - F) \\ \text{inbreeding coefficient } F = 1 - \frac{H}{2pq} \end{array}}$$

Sample frequencies

Sample counts in a random sample of n individuals

multinomial model: $(n_{AA}, n_{Aa}, n_{aa}) \in \mathrm{Mn}(n; D, H, R)$

Genotype frequencies and estimated standard errors

$\hat{D} = \frac{n_{AA}}{n}$, $\hat{H} = \frac{n_{Aa}}{n}$, $\hat{R} = \frac{n_{aa}}{n}$

$s_{\hat{D}} = \sqrt{\frac{\hat{D}(1-\hat{D})}{n-1}}$, $s_{\hat{H}} = \sqrt{\frac{\hat{H}(1-\hat{H})}{n-1}}$, $s_{\hat{R}} = \sqrt{\frac{\hat{R}(1-\hat{R})}{n-1}}$

Allele frequencies

$\hat{p} = \frac{2n_{AA}+n_{Aa}}{2n}$, $\hat{q} = \frac{2n_{aa}+n_{Aa}}{2n}$

$\mathrm{Var}(\hat{p}) = \frac{pq}{2n}(1+F)$, $s_{\hat{p}} = s_{\hat{q}} = \sqrt{\frac{\hat{p}\hat{q}}{2n}(1 + \hat{F})}$, $\hat{F} = 1 - \frac{\hat{H}}{2\hat{p}\hat{q}}$

Ex 5: CCR5 gene

Human chemokine receptor gene

two alleles: A = no deletion, $a = \Delta 32$ deletion

genotype aa is resistant to HIV-1

Paris sample: $n = 294$, electrophoresis results

Band A (long)	—	—	
Band a (short)	—		—
Sample counts	64	224	6
Genotype	Aa	AA	aa

$\hat{D} = \frac{224}{294} = 0.76$, $\hat{H} = 0.22$, $\hat{R} = 0.02$

$s_{\hat{D}} = 0.025$, $s_{\hat{H}} = 0.024$, $s_{\hat{R}} = 0.008$

$\hat{p} = 0.87$, $\hat{q} = 0.13$, $\hat{F} = 0.03$, $s_{\hat{p}} = s_{\hat{q}} = 0.014$

Basques sample: $n = 111$, $\hat{q} = 0.018$, $s_{\hat{q}} = 0.009$

population founded 18000 years ago by a few imm.

1.3 Random mating and HWE

Dynamics of population frequencies over generations:

$$(D_0, H_0, R_0) \rightarrow (D_1, H_1, R_1) \rightarrow (D_2, H_2, R_2) \rightarrow \ldots$$

Hardy-Weinberg principle

for given p_0 whatever are (D_0, H_0, R_0) we get

$D_1 = p_0^2$, $H_1 = 2p_0q_0$, $R_1 = q_0^2$, $p_1 = p_0$, $q_1 = q_0$

offspring inherit genes, not genotypes

$$\boxed{\text{H-W Equilibrium: } D = p^2,\ H = 2pq,\ R = q^2}$$

Hardy-Weinberg assumptions

1. diploid organisms
2. non-overlapping generations
3. effectively infinite population size $N = \infty$
4. random mating = panmixia
5. equal allele frequencies in the sexes
6. no mutation, 7. no selection, 8. no migration

Chi-square test of HWE

Test H_0: HWE using statistic $X^2 = \Sigma_{\text{cells}} \frac{(\text{obs}-\text{exp})^2}{\text{exp}}$

$$\boxed{\begin{array}{l} \text{Asymptotic null distribution } X^2 \in \chi^2_{df} \\ \quad df = \text{number of phenotypes} - \text{number of alleles} \\ \quad \text{when df} = 1 \text{ use normal distribution table} \end{array}}$$

Ex 6: RFLP

Expected (under HWE) genotype frequences

$\hat{D}_0 = \hat{p}^2 = 0.375$, $\hat{H}_0 = 2\hat{p}\hat{q} = 0.475$, $\hat{R}_0 = \hat{q}^2 = 0.150$

Cells	AA	Aa	aa	Total
Observed counts	88	130	32	$n = 250$
Expected counts	93.6	118.7	37.6	$n = 250$
$(\text{obs} - \text{exp})^2/\text{exp}$	0.335	1.076	0.834	$X^2 = 2.25$

P-value of the test: since df $= 3 - 2 = 1$

$P(X^2 \geq 2.25) = P(|\sqrt{X^2}| \geq 1.5)$

$\approx 2(1 - \Phi(1.5)) = 0.134$, accept H_0

Chi-square test and inbreeding coefficient: $X^2 = n \cdot \hat{F}^2$

Ex 5: CCR5 gene

Paris sample

$$X^2 = 294 \cdot (0.03)^2 = 0.26, \text{ df} = 1, \text{ accept HWE}$$

Estimation under HWE

Single gene recessive disease:

two phenotypes and two alleles, df $= 2 - 2 = 0$

cannot test HWE from phenotypes

Assuming HWE use estimate $\hat{q} = \sqrt{\bar{R}}$ with $s_{\hat{q}} = \sqrt{\frac{1-R}{4n}}$

Ex 7: cystic fibrosis

CFTR gene, two alleles: normal A, mutant a

aa causes a severe condition, Caucasian $R = \frac{1}{2500}$

Assuming HWE for Caucasians

$$q = \sqrt{R} = 0.02 \text{ and } H = 2 \cdot 0.02 \cdot 0.98 = \frac{1}{26}$$

Carriers to affected ratio $\frac{H}{R} = \frac{2p}{q} \approx \frac{2}{q}$

Propagation of error method

$$f(\hat{R}) \approx f(R) + f'(R)(\hat{R} - R) + \tfrac{1}{2}f''(R)(\hat{R} - R)^2$$
$$\mathrm{E}(f(\hat{R})) \approx f(R) + \tfrac{1}{2}f''(R)\,\mathrm{Var}(\hat{R})$$

If $f(x) = \sqrt{x}$, then $\mathrm{E}(\hat{q}) = \mathrm{E}(\sqrt{\bar{R}}) \approx \sqrt{R} - \tfrac{1}{8}R^{-3/2}\frac{R(1-R)}{n}$

$$\mathrm{Var}(\hat{q}) = \mathrm{E}(\hat{R}) - (\mathrm{E}(\hat{q}))^2 \approx \frac{1-R}{4n}$$

Literature:
1. D.L.Hartl, A.G.Clarc. Principle of population genetics. Sinauer Associates, 2007.
2. R.Nielson, M. Statkin. An introduction to population genetics: theory and applications, Sinauer Associates. 2013.