

Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP



Godfrey Harold **Hardy** (1877-1947)

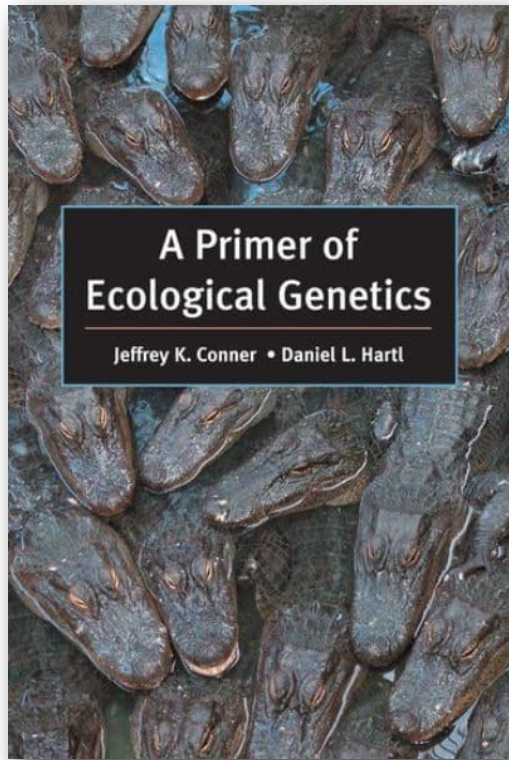


Wilhelm **Weinberg** (1862-1937)

G. H. Hardy: "Mendelian proportions in a mixed population". Science 28, 1908

Wilhelm Weinberg: "Über den Nachweis der Vererbung beim Menschen", Jahrbefte des Vereines für Vater. Naturkunde in Württemberg, 1908

W. E. Castle: "The laws of Galton and Mendel and some laws governing race improvement by selection", Proc. Amer. Acad. Arts Sci.. 35, 1903



A Primer of Ecological Genetics



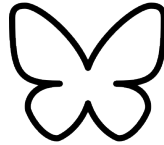
Chapter 2 - Hardy-Weinberg - Page 25-36

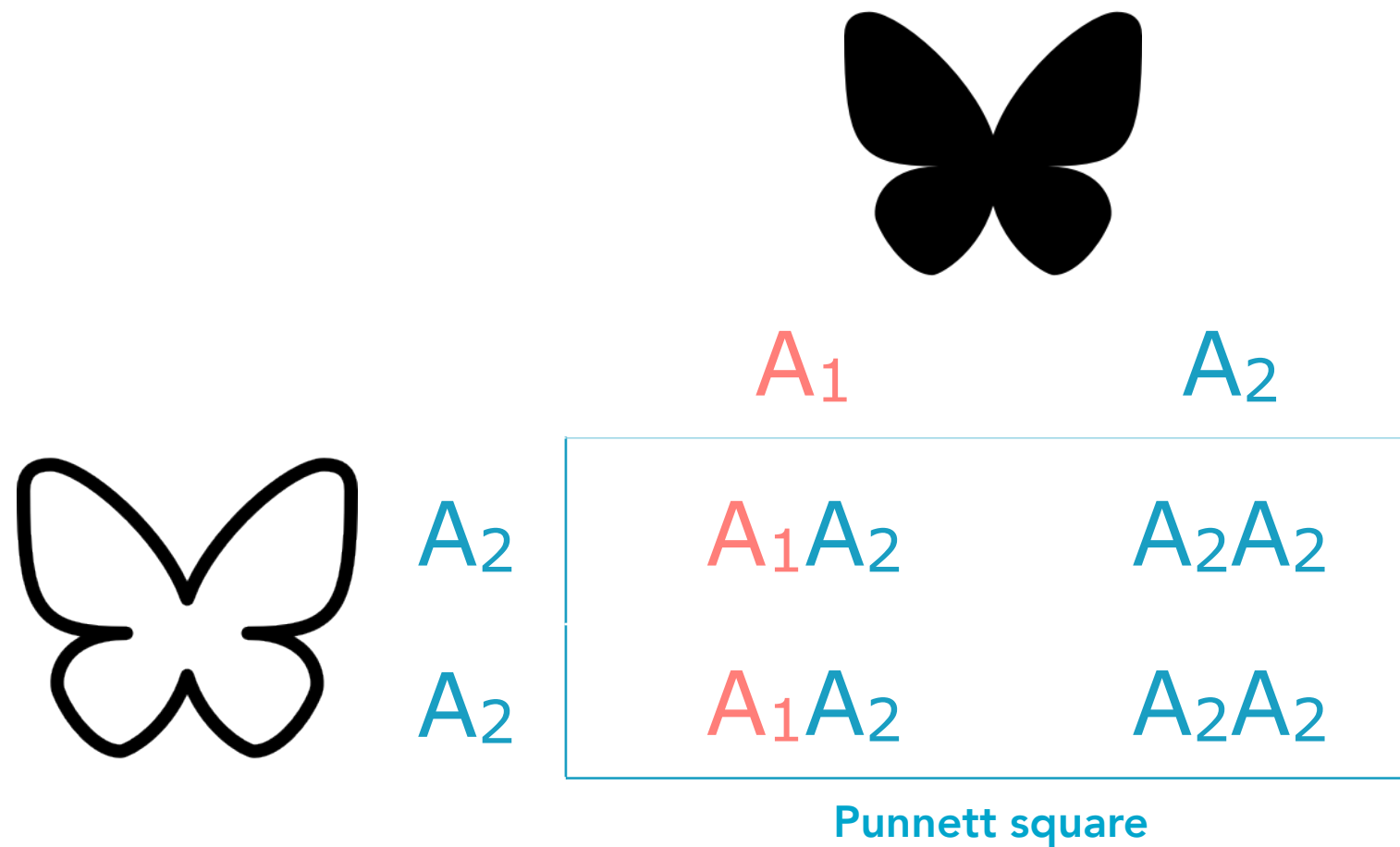
Simple model of population genetics:

- ▶ Random Mating
- ▶ No Mutation
- ▶ Large Population Size → No/Low Genetic Drift
- ▶ No Natural Selection
- ▶ No Immigration

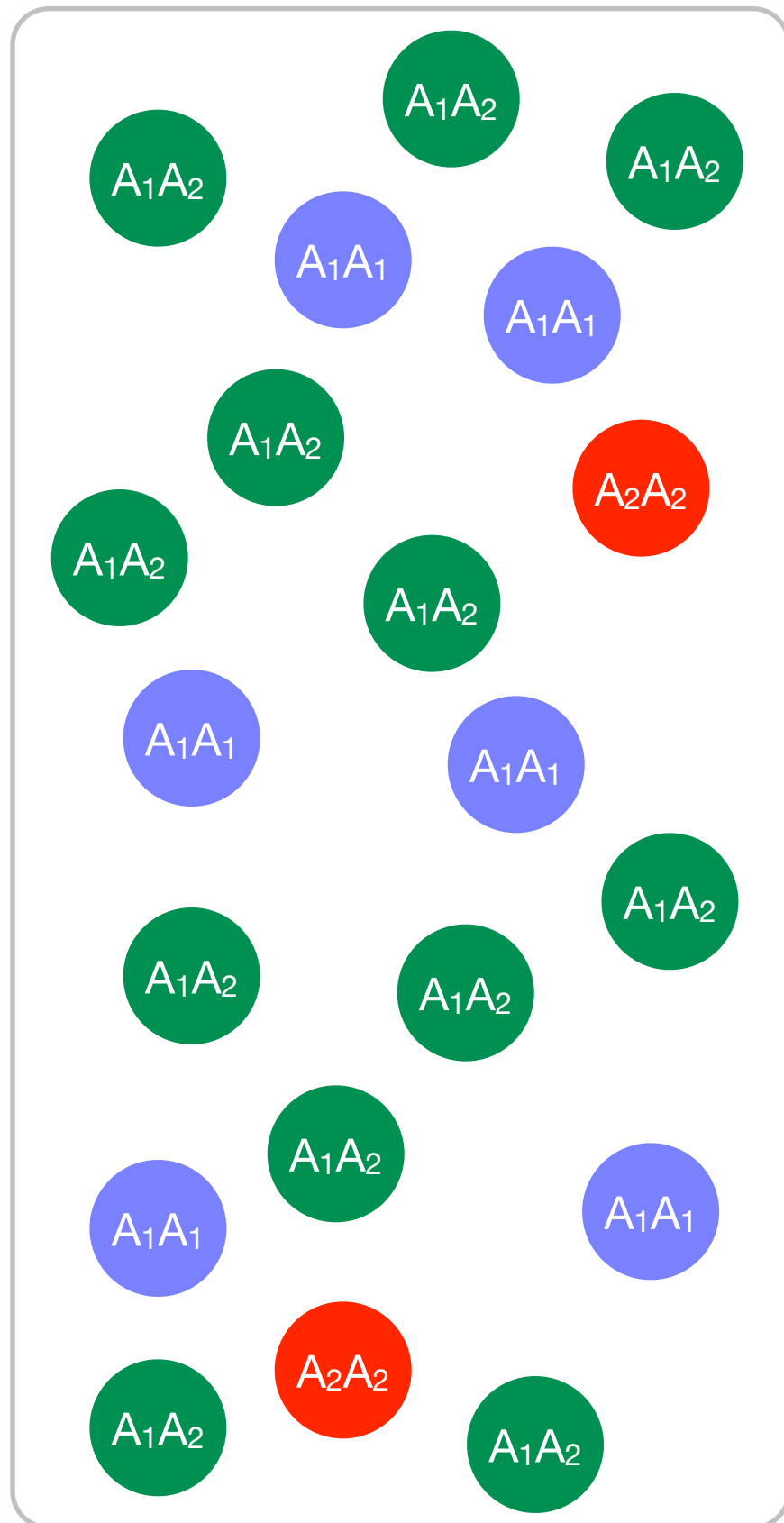


PopGen ▷ Hardy-Weinberg Principle

			
Phenotype	dark	dark	light
	Homozygote	Heterozygote	Homozygote
Genotype	AA	Aa	aa
	AA	AB	BB
	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂



$$A_1A_2 \times A_2A_2 \rightarrow 50\% A_1A_2 \text{ and } 50\% A_2A_2$$



Genotype Frequency

$$N_{A_1A_1} = 6$$

$$x_{11} = \frac{6}{20} = 0.3 \cong P$$

$$N_{A_2A_2} = 2$$

$$x_{22} = \frac{2}{20} = 0.1 \cong Q$$

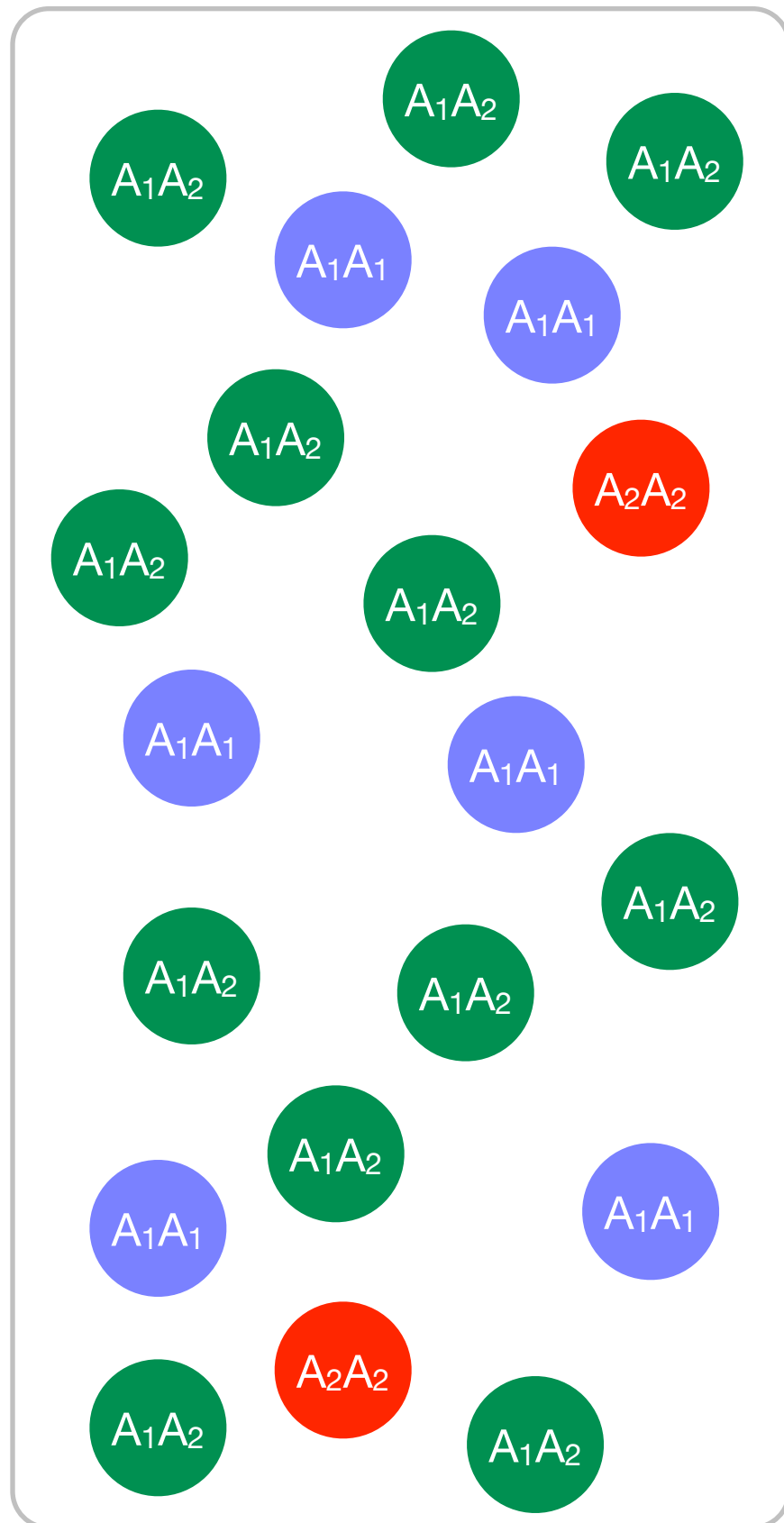
$$N_{A_1A_2} = 12$$

$$x_{12} = \frac{12}{20} = 0.6 \cong H$$

$$N_{Total} = 20$$

$$x_{11} + x_{22} + x_{12} = 1.0$$

$$P + Q + H = 1.0$$



Allele Frequency

$$N_{A_1} = 24$$

$$N_{A_2} = 16$$

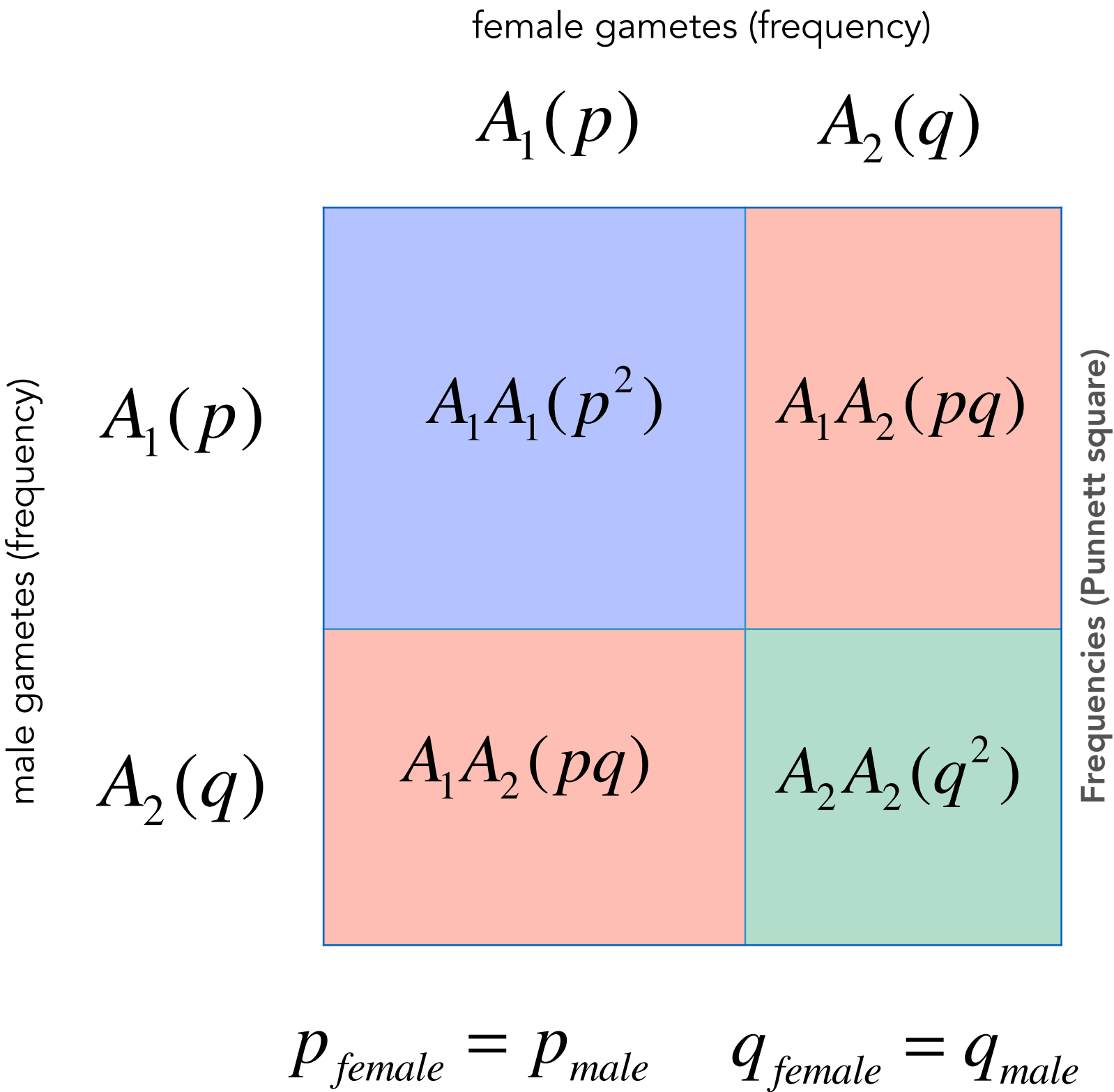
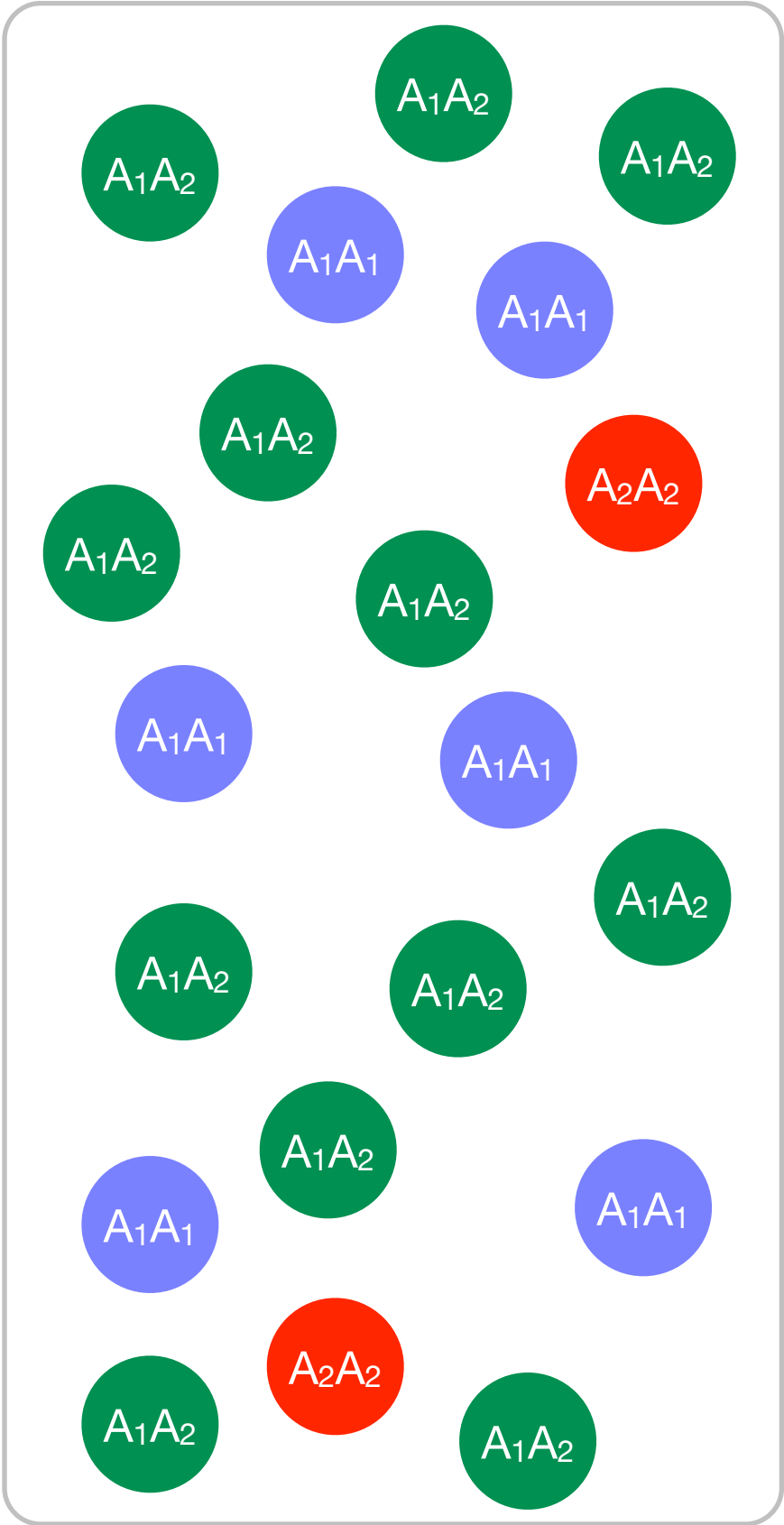
$$N_{Total} = 40$$

$$f(A_1) = \frac{24}{40} = 0.6 = p$$

$$f(A_2) = \frac{16}{40} = 0.4 = q$$

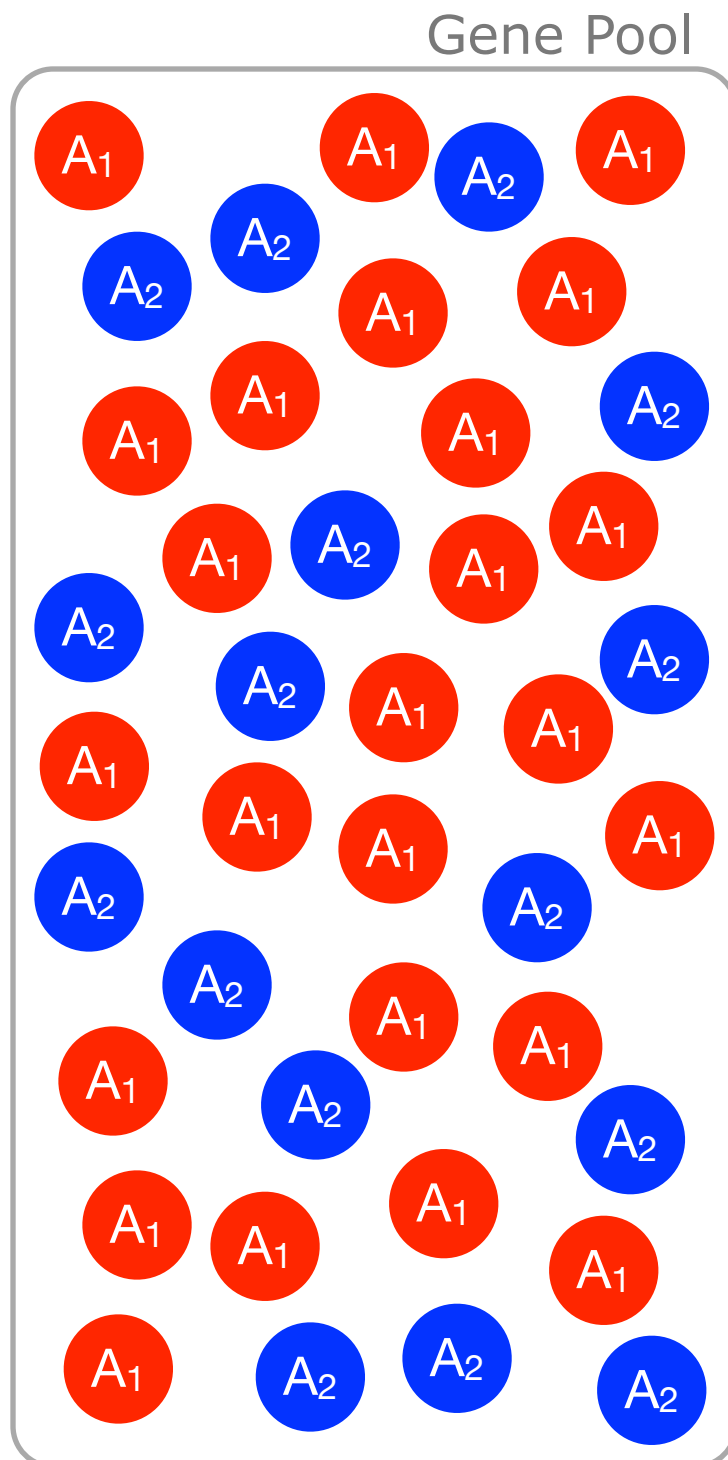
$$f(A_1) + f(A_2) = 1.0$$

$$p + q = 1.0$$



PopGen ► Hardy-Weinberg Principle

The **Hardy-Weinberg law** describes the equilibrium state of a single locus in a randomly mating diploid population that is free of other evolutionary forces, such as mutation, migration, and genetic drift.



$$\text{A}_1 = 24 \Rightarrow f(\text{A}_1) = p = \frac{24}{40} = 0.6$$

$$\text{A}_2 = 16 \Rightarrow f(\text{A}_2) = q = \frac{16}{40} = 0.4$$

$$\text{A}_1 \text{ A}_1 = p^2 = 0.36$$

$$\text{A}_2 \text{ A}_2 = q^2 = 0.16$$

$$\left. \begin{array}{l} \text{A}_1 \text{ A}_2 = pq = 0.24 \\ \text{A}_2 \text{ A}_1 = qp = 0.24 \end{array} \right\} = 2pq = 0.48$$

$$p^2 + 2pq + q^2 = 0.36 + 0.48 + 0.16 = 1$$

Genotype frequencies

$$\text{freq}(A_1A_1) = \frac{N_{11}}{N} = p^2 \cong P$$

$$\text{freq}(A_1A_2) = \frac{N_{12}}{N} = 2pq \cong H$$

$$\text{freq}(A_2A_2) = \frac{N_{22}}{N} = q^2 \cong Q$$

$$p^2 + 2pq + q^2 = (p + q)^2 = 1$$

Allele frequencies

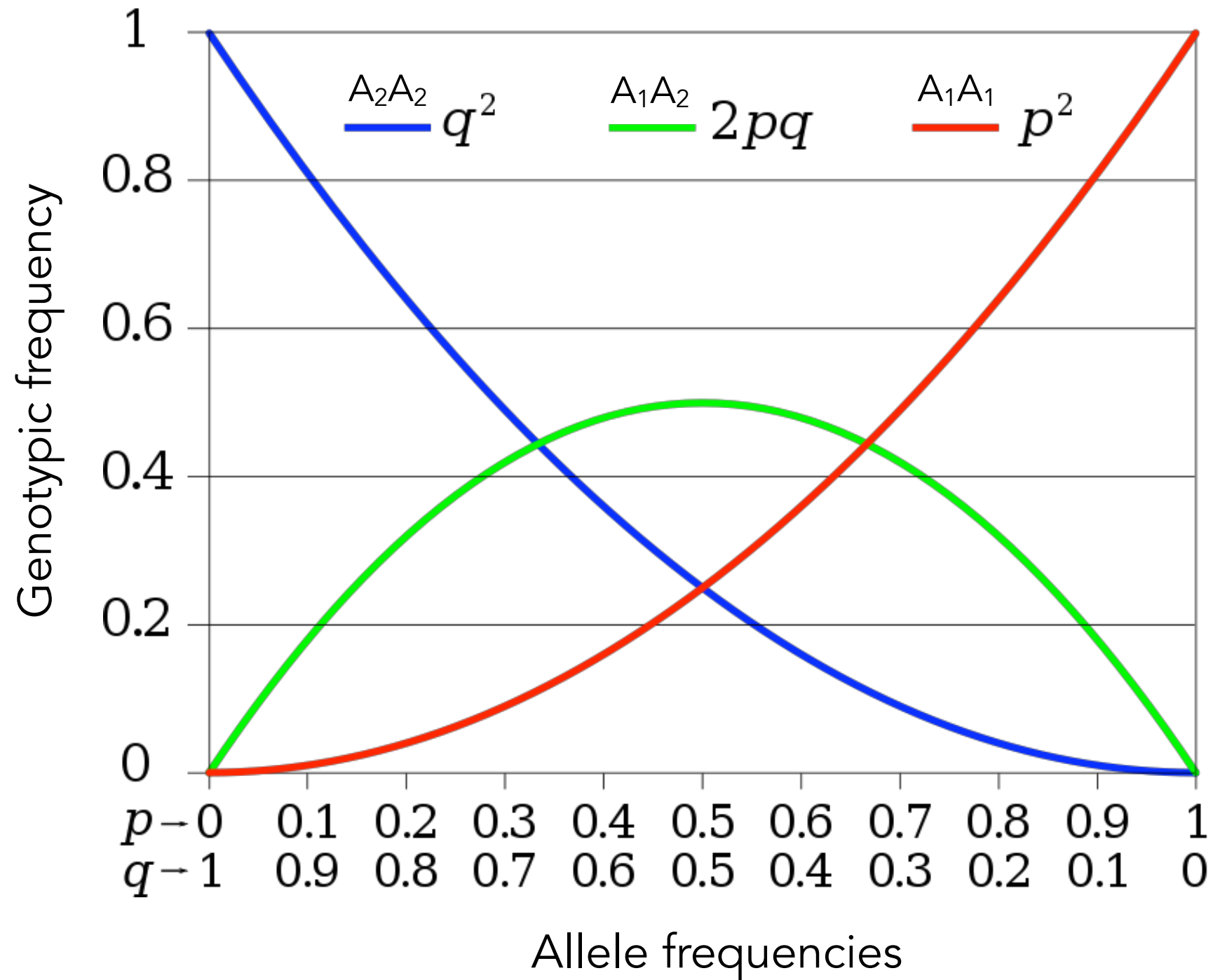
$$\text{freq}(A_1) = p = \frac{(2N_{11} + N_{12})}{2N}$$

$$\text{freq}(A_2) = q = \frac{(2N_{22} + N_{12})}{2N}$$

$$p + q = 1$$

	A_1A_1	A_1A_2	A_2A_2	
	p^2	$2pq$	q^2	$q^2+2pq+q^2$
$p=0.10$	0.010	0.180	0.810	1.000
$p=0.25$	0.063	0.375	0.563	1.000
$p=q=0.5$	0.250	0.500	0.250	1.000
$p=0.75$	0.563	0.375	0.063	1.000
$p=0.90$	0.810	0.180	0.010	1.000

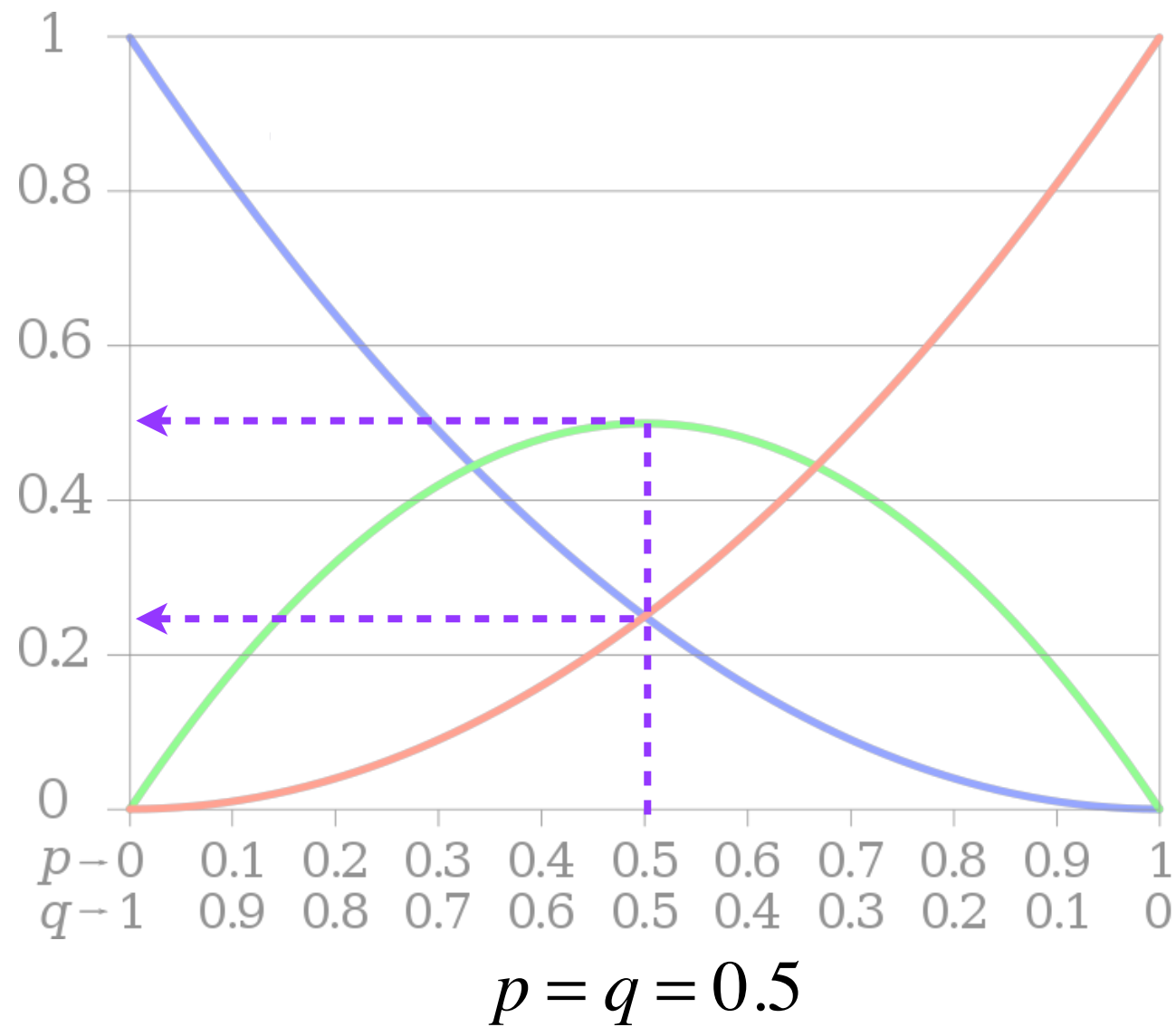
PopGen ▷ Hardy-Weinberg Principle



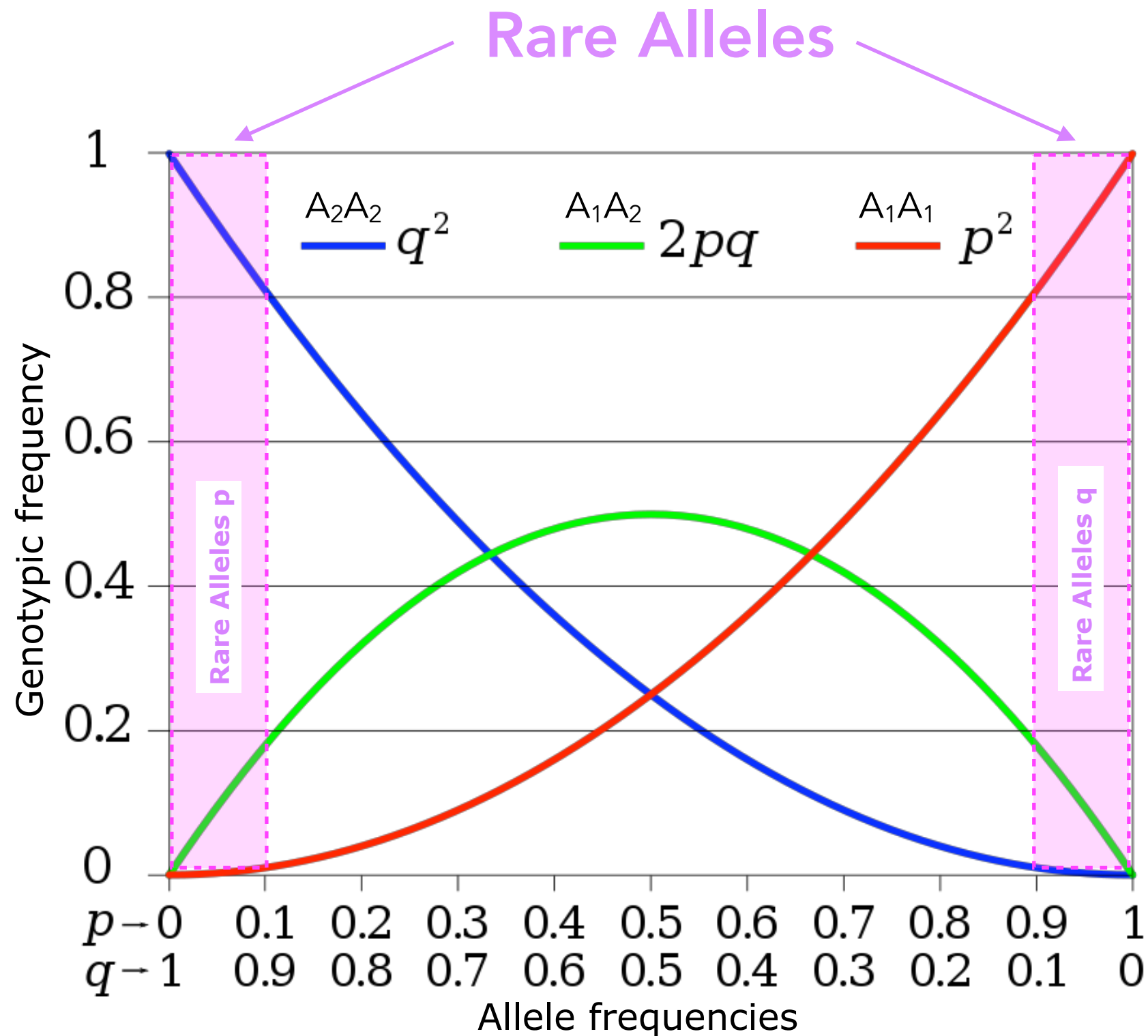
PopGen ▷ Hardy-Weinberg Principle

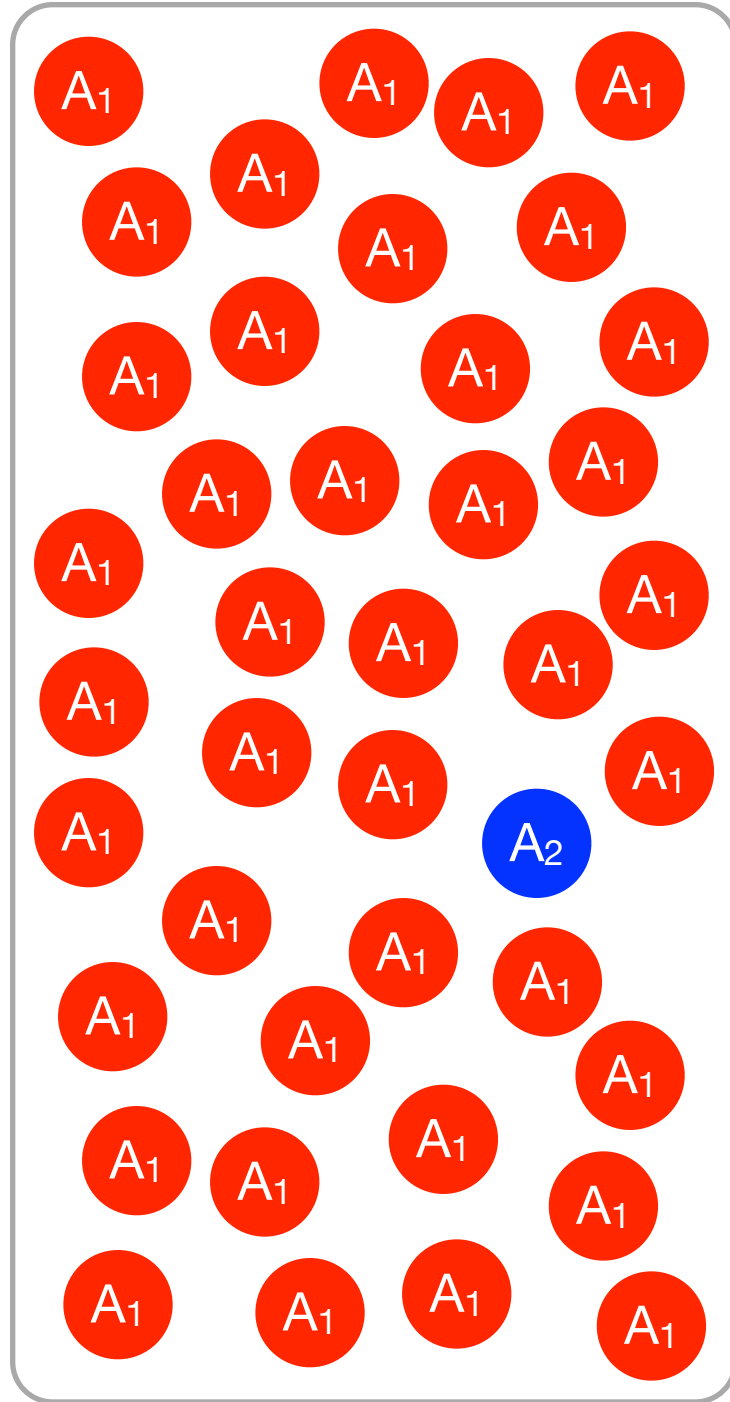
Heterozygote: 50%

Homozygote: 2x25%



PopGen ▷ Hardy-Weinberg Principle





One of the consequences of the Hardy-Weinberg law concerns the genotype occupied by **rare alleles**.

$$A_1 = 39 \Rightarrow f(A_1) = p = 0.975$$

$$A_2 = 1 \Rightarrow f(A_2) = q = 0.025$$

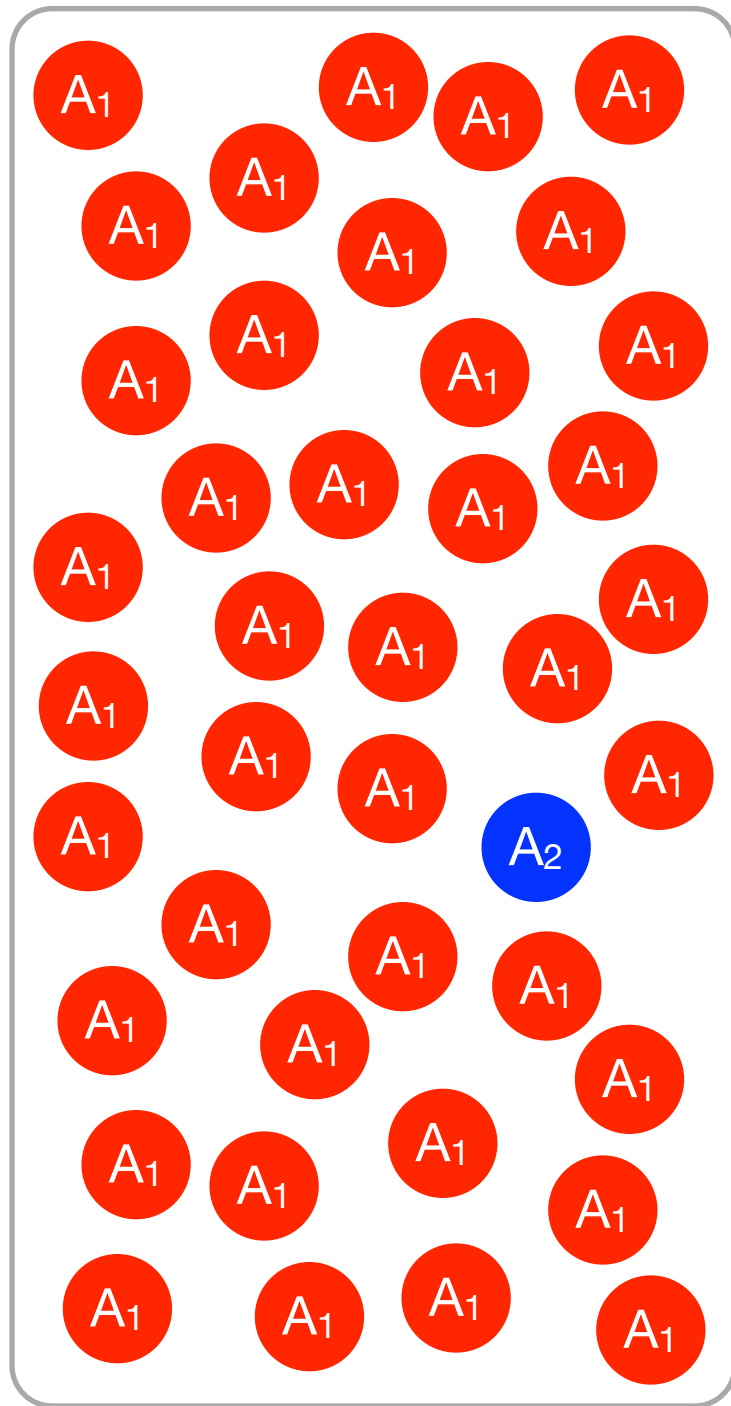
2.5%

$$A_1 A_1 = p^2 = 0.95$$

$$A_2 A_2 = q^2 = 0.000625$$

$$\left. \begin{array}{cc} A_1 & A_2 \\ A_2 & A_1 \end{array} \right\} = 2qp = 0.04875$$

5%



Rare alleles are especially susceptible to loss during a bottleneck. However, the loss of rare alleles will have little effect on the **persistence of heterozygosity**.

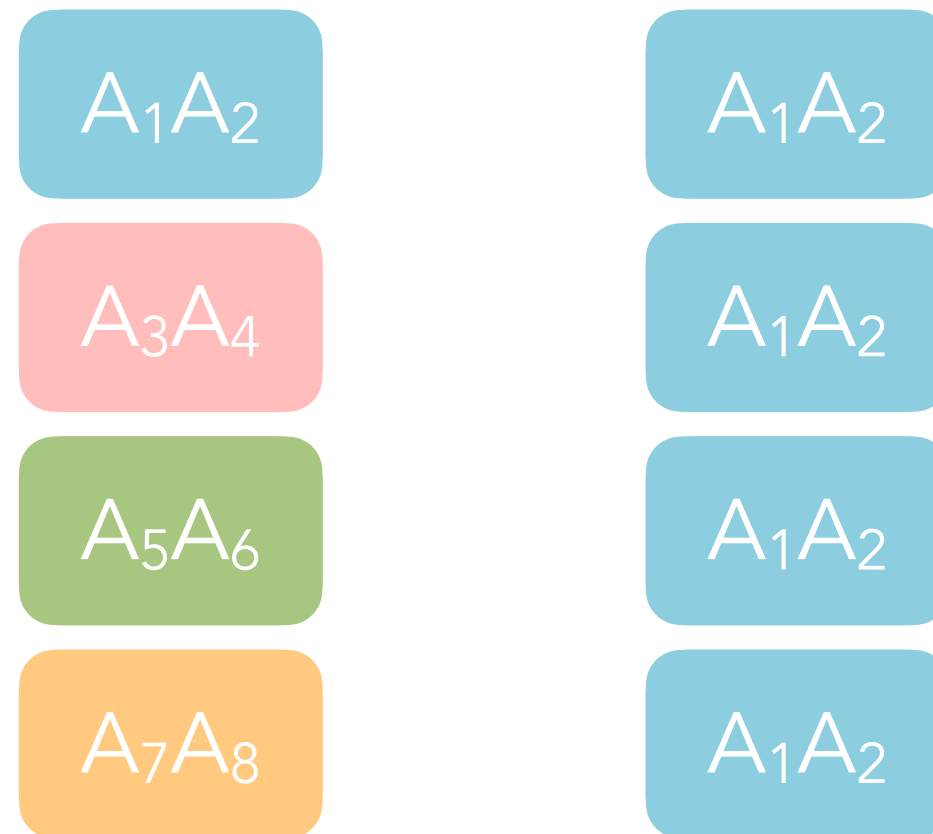
$$\frac{A_1A_2}{A_1A_1} = \frac{2pq}{p^2} = \frac{2q}{p} \xrightarrow{p \approx 1} 2q$$



Founder events are known to **decrease the genetic diversity** of the population, and are often followed by a **demographic expansion**. It has been shown, both theoretically, and empirically, that **allelic richness is more sensitive than heterozygosity to founder events followed by expansions**, since allelic richness does not consider abundances of the alleles but only their presence (a rare allele that is lost in a founder event will probably not affect heterozygosity much, but the loss does reduce allelic richness). Moreover, **allelic richness is more indicative of the evolutionary potential of the population in the long-run** in these scenarios, as the existence of alleles, rather than their frequencies, holds a significant part of potential for response to selection, as **selection limits are determined by the initial allelic composition rather than by levels of heterozygosity**.



Source: Greenbaum et al. (2015) PLoS ONE



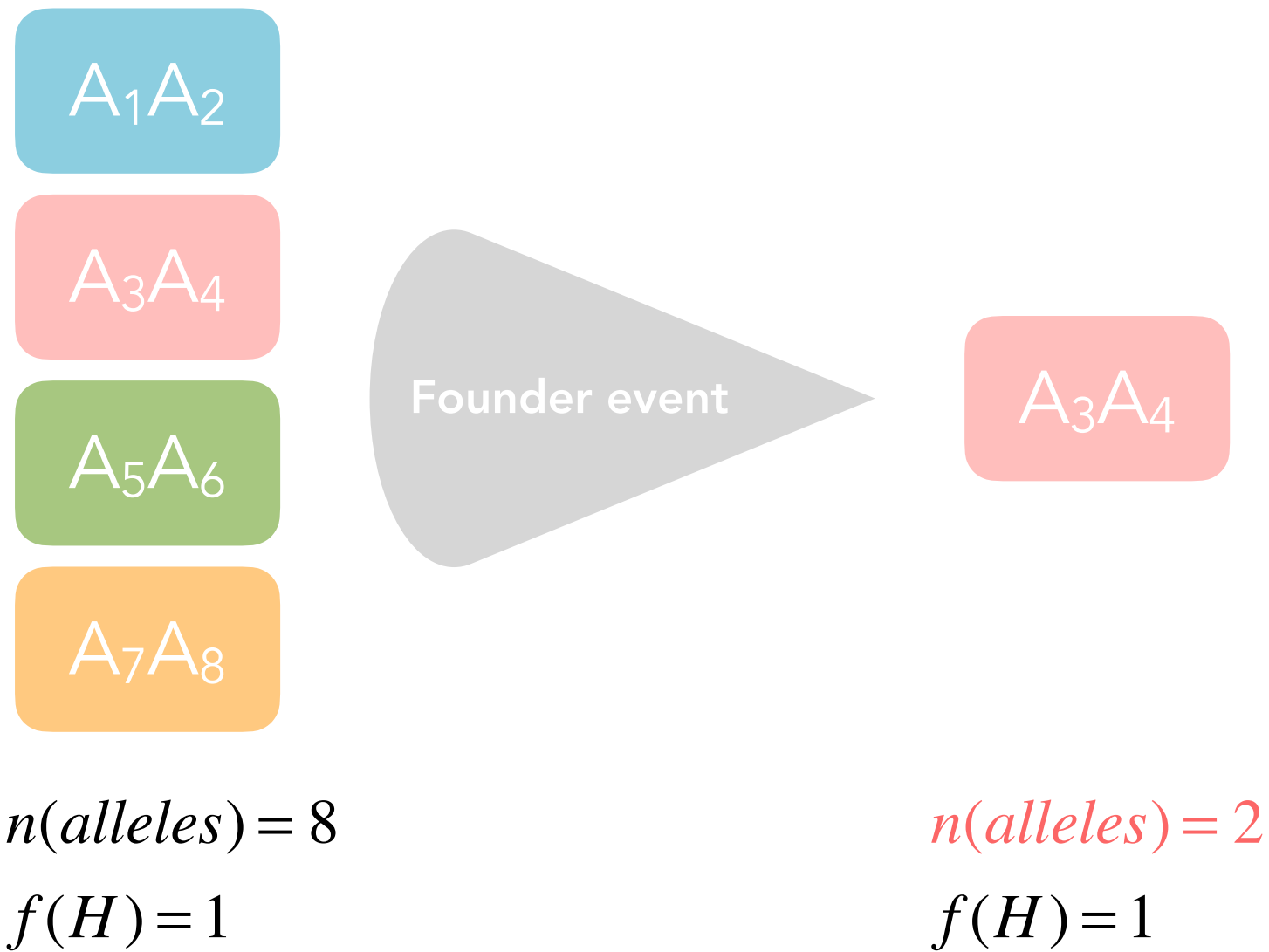
$$n(\text{alleles}) = 8$$

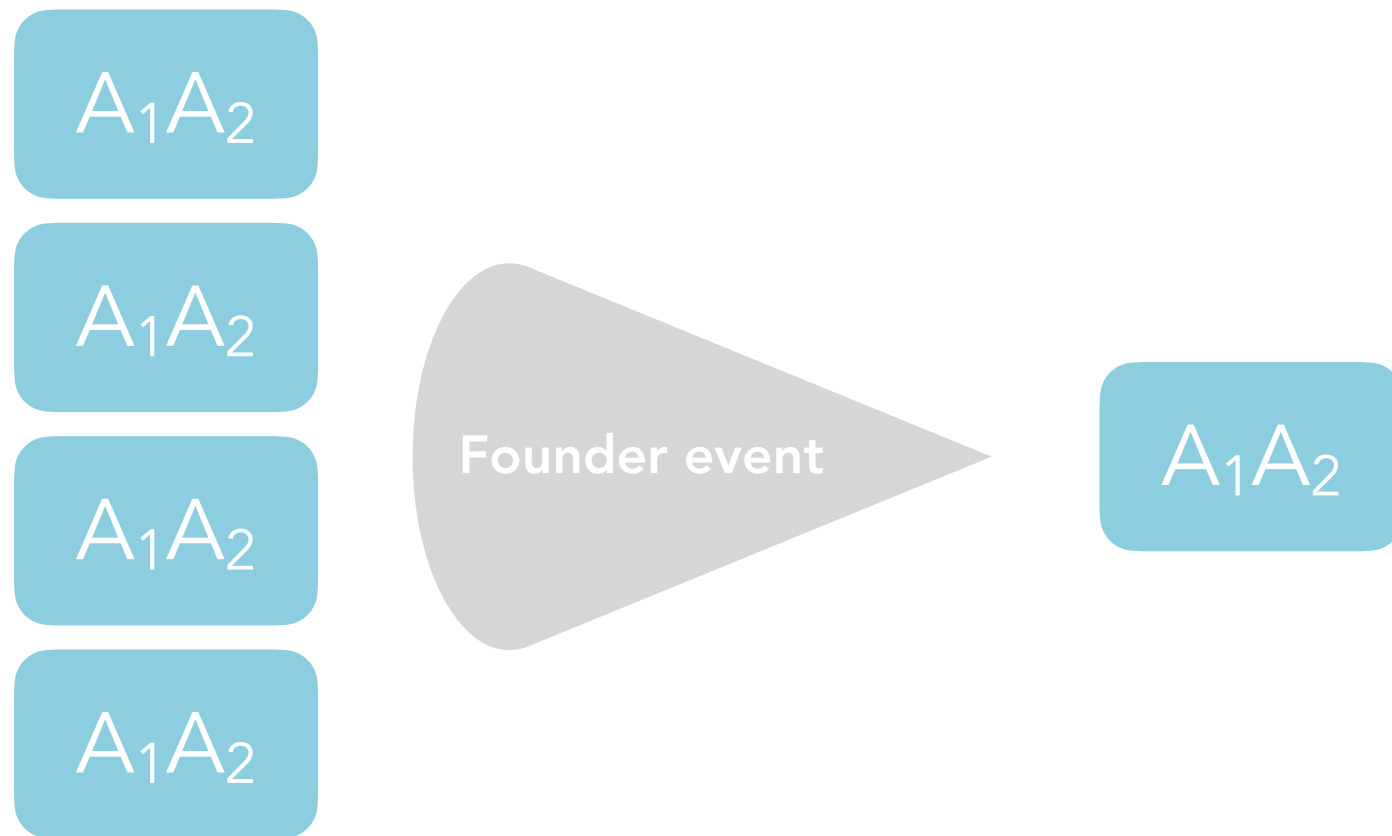
$$f(H) = 1$$

$$n(\text{alleles}) = 2$$

$$f(H) = 1$$

PopGen ▷ Hardy-Weinberg Principle





$$n(\text{alleles}) = 2$$

$$f(H) = 1$$

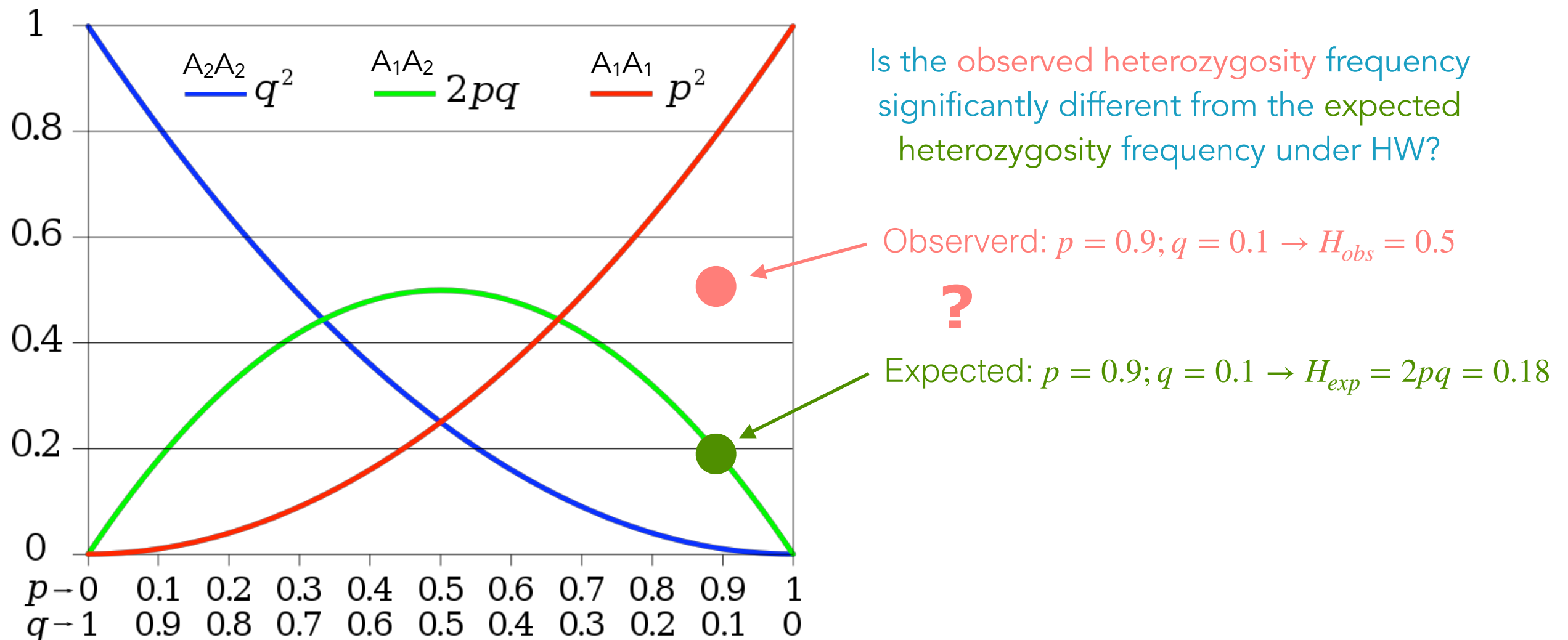
$$n(\text{alleles}) = 2$$

$$f(H) = 1$$

The Hardy-Weinberg theorem only works if the following assumptions for the population are met:

- ▶ diploid individuals
- ▶ reproduces by sexual means
- ▶ mating at random
- ▶ population is infinite
- ▶ no mutation
- ▶ no (natural) selection
- ▶ no genetic exchange with other populations

PopGen ▷ Hardy-Weinberg Principle



Pearson's chi-squared test can be used to determine whether there is a statistically significant difference between the **expected frequencies** and the **observed frequencies** in one or more categories of a contingency table.

The **chi-square statistic** is a mathematical tool used in statistics to assess the independence or association between categorical variables. It measures the extent to which observed data differ from what would be expected if the variables were independent. It is often used in **hypothesis testing**, where researchers compare observed data with expected data to determine if there is a significant relationship between variables.

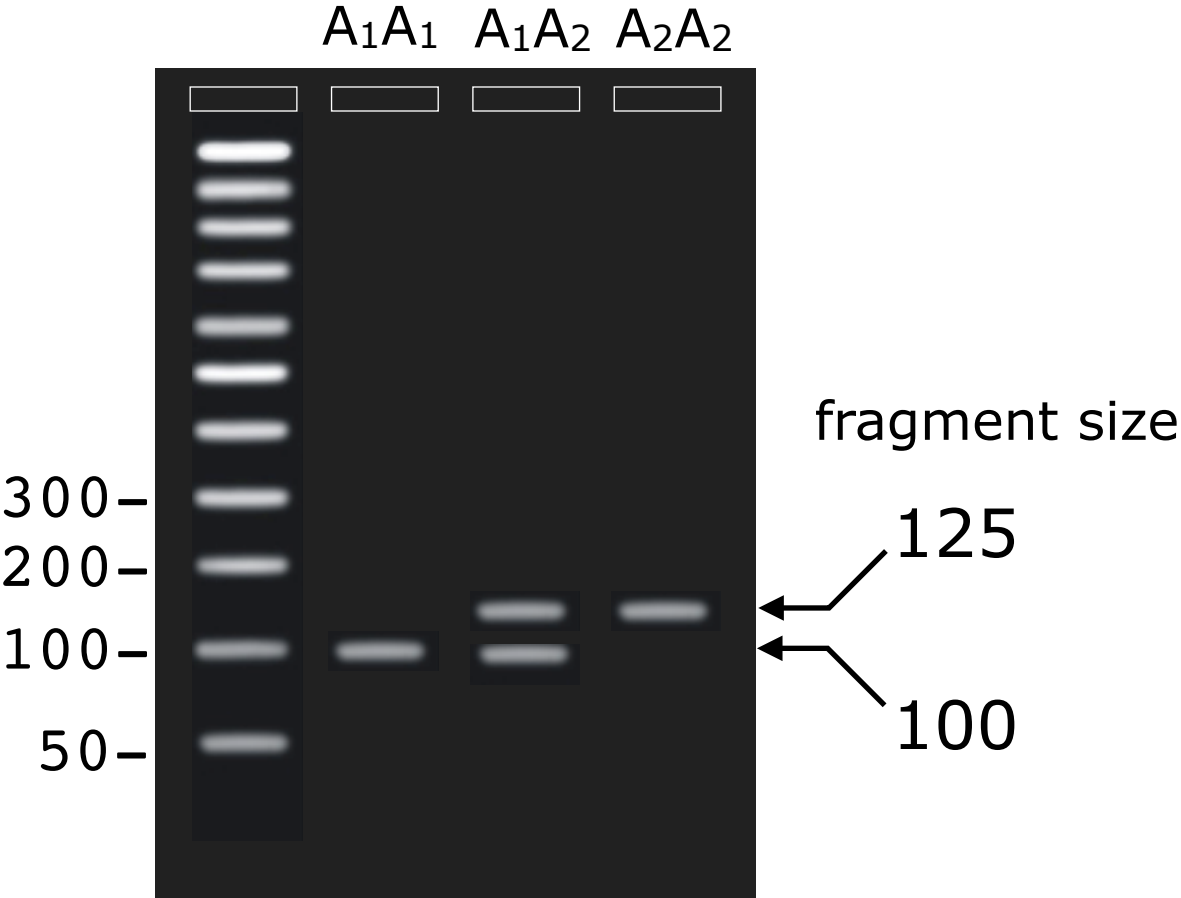
The formula for the chi-square statistic involves summing the squared differences between the observed and expected frequencies, which are then standardised and compared to a chi-squared distribution to determine statistical significance. If the calculated chi-square value is sufficiently different from what would be expected by chance, it suggests a significant relationship between the variables.

$$\chi^2 = \sum \frac{(\textit{observed} - \textit{expected})^2}{\textit{expected}}$$

Observed:

	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
	100/100	100/125	125/125	total
	5	12	6	23

Genotype



PopGen ▷ Hardy-Weinberg Principle

Punnett square	freq(A ₁) = p _f	freq(A ₂) = q _f
freq(A ₁) = p _m	freq(A ₁ A ₁) = p _f × p _m	freq(A ₂ A ₁) = q _f × p _m
freq(A ₂) = q _m	freq(A ₁ A ₂) = p _f × q _m	freq(A ₂ A ₂) = q _f × q _m

$$freq(A_1A_1) = p_f \cdot p_m \xrightarrow{p_m=p_f} p^2$$

$$freq(A_1A_2) = 2(p_f \cdot p_m) \xrightarrow[q_m=q_f]{p_m=p_f} 2pq$$

$$freq(A_2A_2) = q_f \cdot q_m \xrightarrow{q_m=q_f} q^2$$

$$p^2 + 2pq + q^2 = (p + q)^2 = 1$$

PopGen ▷ Hardy-Weinberg Principle

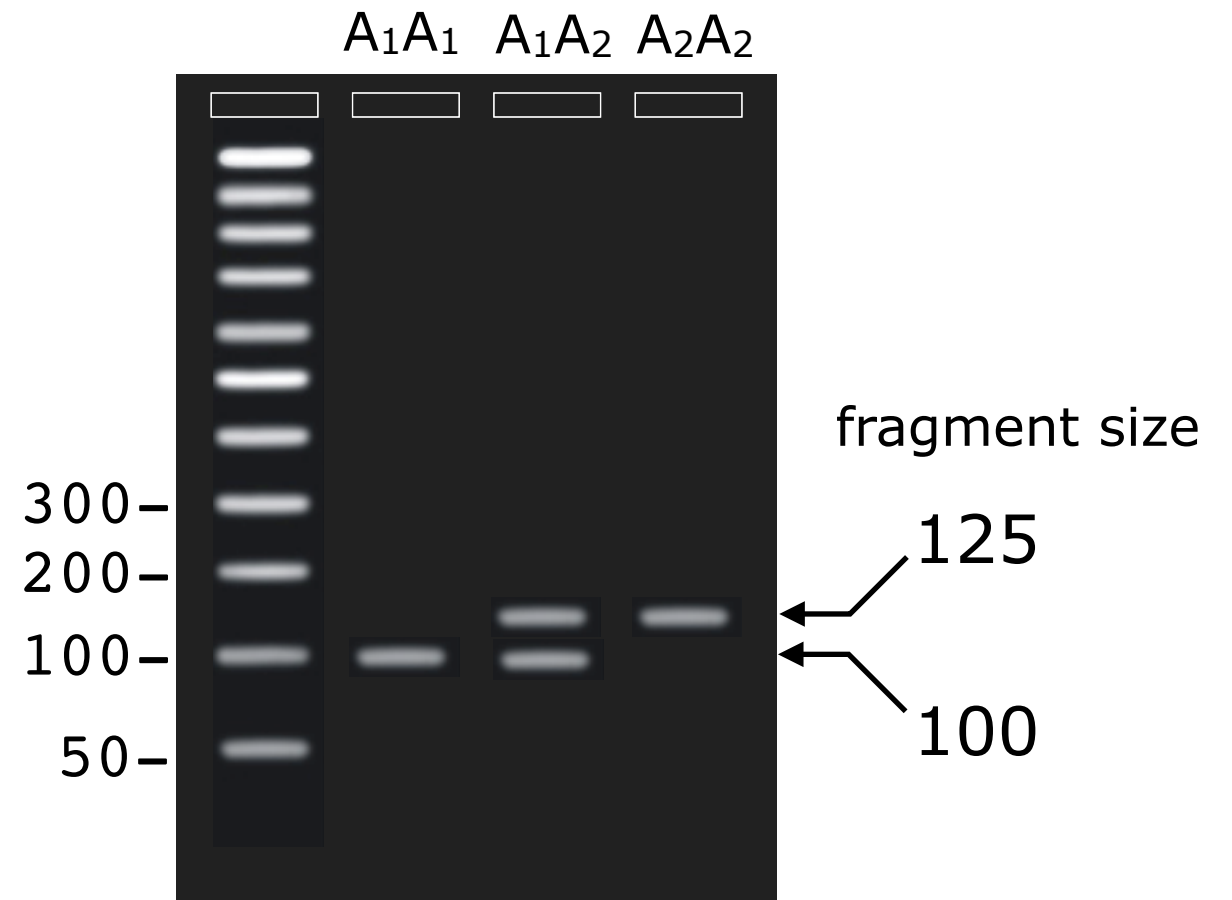
Observed:

	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
	100/100	100/125	125/125	total
	5	12	6	23

$$\hat{p} = \frac{2N_{A_1A_1} + N_{A_1A_2}}{2N_{total}} = \frac{10 + 12}{46} = 0.478$$

$$\hat{q} = \frac{2N_{A_2A_2} + N_{A_1A_2}}{2N_{total}} = \frac{12 + 12}{46} = 0.522$$

Genotype



Observed:

	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
	100/100	100/125	125/125	total
	5	12	6	23

$$\hat{p} = \frac{2N_{A_1A_1} + N_{A_1A_2}}{2N_{total}} = \frac{10 + 12}{46} = 0.478$$

$$\hat{q} = \frac{2N_{A_2A_2} + N_{A_1A_2}}{2N_{total}} = \frac{12 + 12}{46} = 0.522$$

Expected:

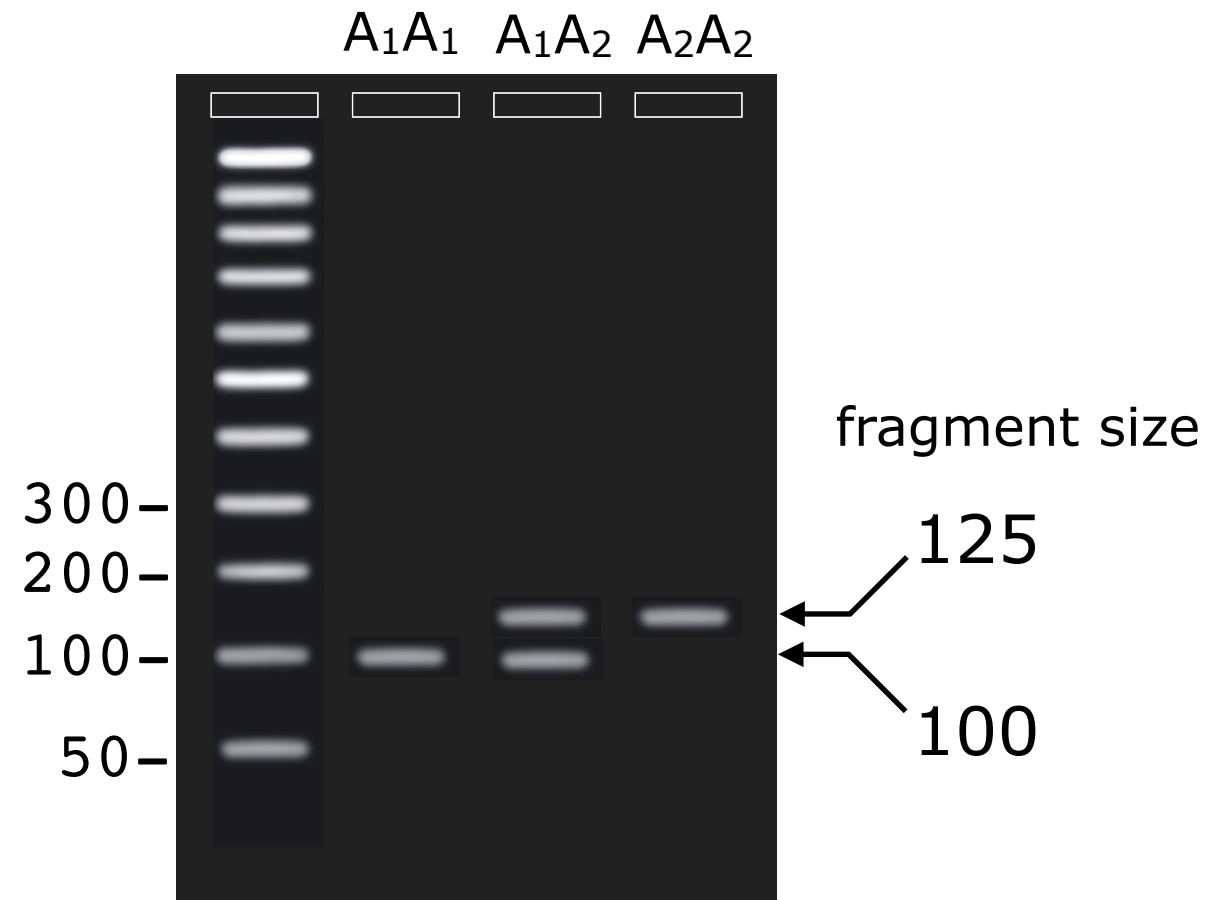
	100/100	100/125	125/125	total
	5.3	11.5	6.3	23.1

$$N_{A_1A_1} = \hat{p}^2 N_{total} = 0.478^2 \cdot 23 = 5.3$$

$$N_{A_1A_2} = 2\hat{p}\hat{q}N_{total} = 2 \cdot 0.478 \cdot 0.522 \cdot 23 = 11.5$$

$$N_{A_2A_2} = \hat{q}^2 N_{total} = 0.522^2 \cdot 23 = 6.3$$

Genotype



PopGen ► Hardy-Weinberg Principle

Observed:

A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
100/100	100/125	125/125	total
5	12	6	23

Expected:

100/100	100/125	125/125	total
5.3	11.5	6.3	23.1

The **chi-square method** provides a statistical test to determine whether the deviation between the observed genotypic and expected Hardy-Weinberg proportions is greater than we would expect by chance alone.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

PopGen ▷ Hardy-Weinberg Principle

Observed:

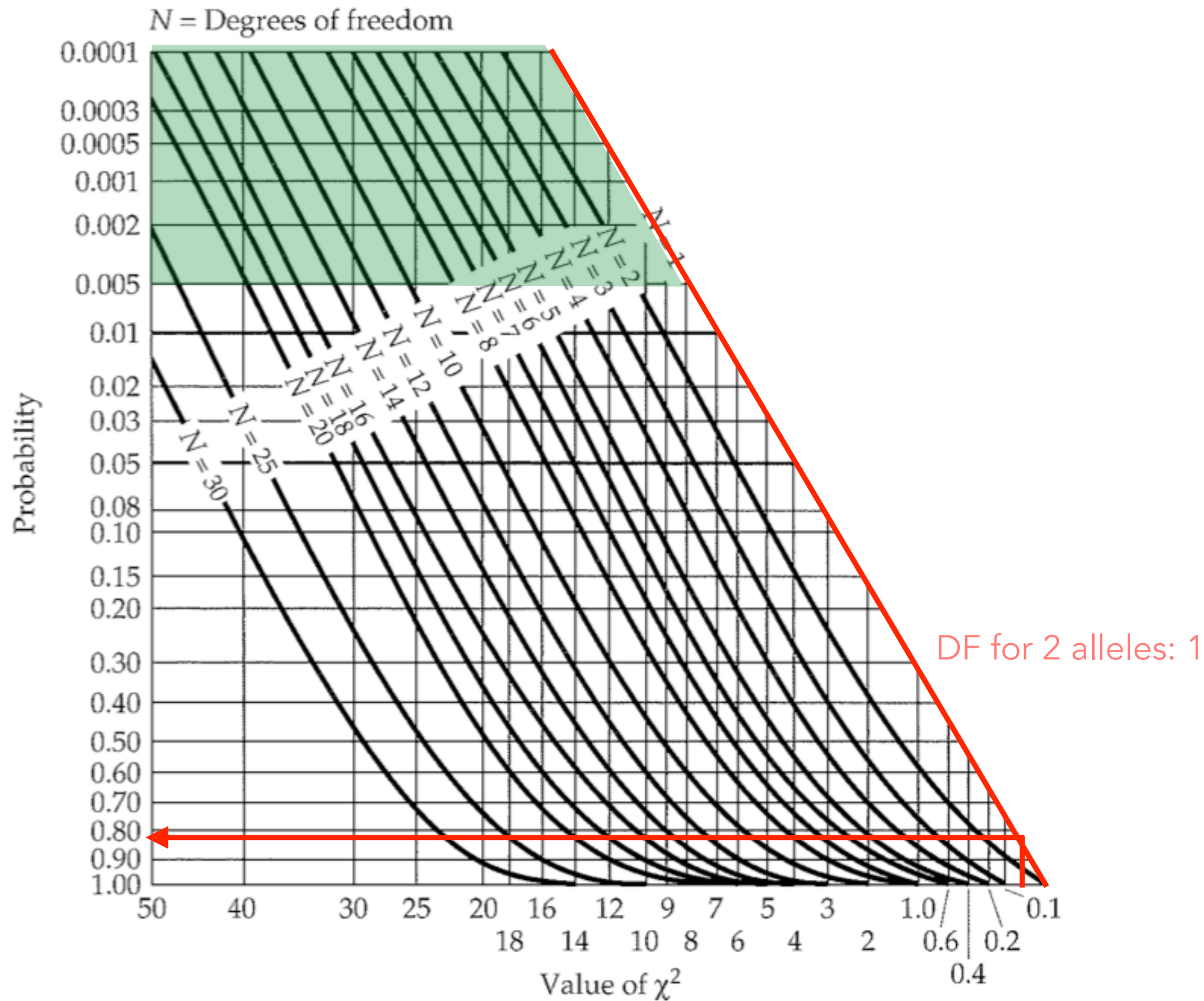
A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
100/100	100/125	125/125	total
5	12	6	23

Expected:

100/100	100/125	125/125	total
5.3	11.5	6.3	23.1

$$\chi^2 = \frac{(5 - 5.3)^2}{5.3} + \frac{(12 - 11.5)^2}{11.5} + \frac{(6 - 6.3)^2}{6.3} = 0.05$$

PopGen ▷ Hardy-Weinberg Principle



PopGen ▷ Hardy-Weinberg Principle


Observed:

A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
100/100	100/125	125/125	total
5	12	6	23

Expected:

100/100	100/125	125/125	total
5.3	11.5	6.3	23.1

$$\chi^2 = \frac{(5 - 5.3)^2}{5.3} + \frac{(12 - 11.5)^2}{11.5} + \frac{(6 - 6.3)^2}{6.3} = 0.05$$

 `1-pchisq(0.05, 1) = 0.82 > 0.05`

H₀ (null hypothesis): The population is in Hardy-Weinberg proportions at this locus.



```
# Genotype Counts
x <- c(AA = 5, AB = 12, BB = 6)
# HW-Test
HWres <- HardyWeinberg::HWChisq(x, c = 0, verbose = TRUE)

##  $\chi^2$  : value of the chi-square statistic
> HWres$chisq
# [1] 0.04752066 ( $\chi^2$ )

## p-value of the chi-square test for HWE
> HWres$pval
# [1] 0.8274351 (p-value)    ← 1-pchisq(0.05, 1)

## Half the deviation from HWE for the AB genotype
> HWres$D
# [1] 0.2608696 (D)
```

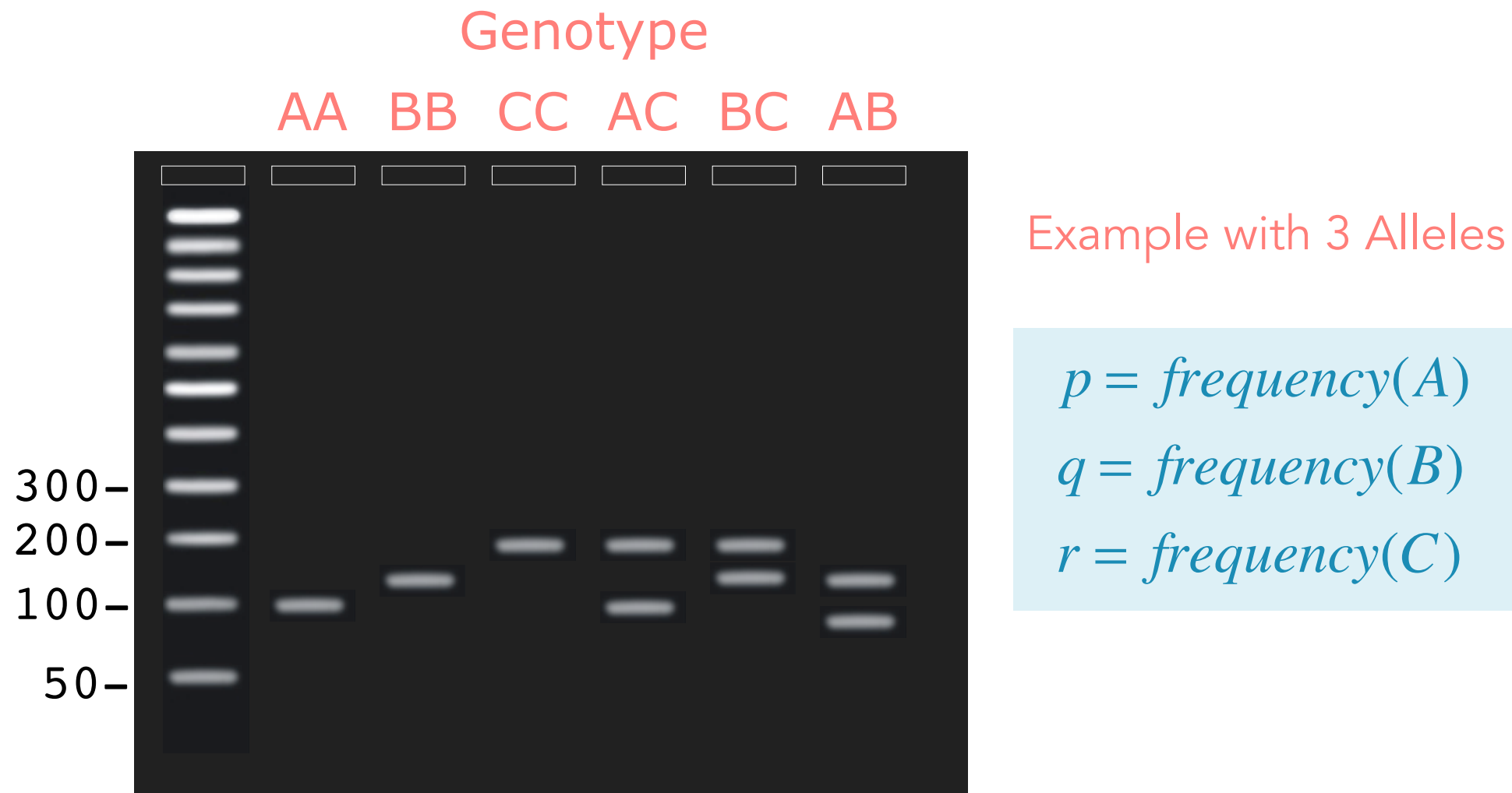

Alternatively, the chi-square statistic may be expressed as $\chi^2 = \frac{D^2}{p^2q^2n}$

where $D = \frac{1}{2}(\textit{observed}_{AB} - \textit{expected}_{AB})$ indicates the deviation

from independence for the heterozygote.

PopGen ▷ Hardy-Weinberg Principle

The **generalisation** of the Hardy-Weinberg law to **multiple alleles** requires no new ideas.



$$(p + q + r)^2 = p^2 + 2pq + q^2 + 2pr + 2qr + r^2$$

The χ^2 test can be extended from 2-alleles to multiple alleles.

$$\chi^2 = \sum_{i=1}^k \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

2 alleles

$$\chi^2 = \frac{(N_{11} - p^2 N)^2}{p^2 N} + \frac{(N_{12} - 2pqN)^2}{2pqN} + \frac{(N_{22} - q^2 N)^2}{q^2 N}$$

n alleles

$$\chi^2 = \sum_i \frac{(N_{ii} - p_i^2 N)^2}{p_i^2 N} + \sum_{i < j} \frac{(N_{ij} - 2p_i p_j N)^2}{2p_i p_j N}$$

PopGen ▷ Hardy-Weinberg Principle

The frequency of the i th allele will be called p_i ($i=1\dots n$). As before, the frequency of the A_iA_j genotype will be called x_{ij} . As with the two-allele case, the sum of all the genotype frequencies must add to one.

$$x_{11} + x_{22} + \dots + x_{nn} + x_{12} + x_{13} + \dots + x_{(n-1)n} = \sum_{i=1}^n \sum_{j \geq 1}^n x_{ij} = 1$$

the frequency of the i th allele is:

$$p_i = x_{ii} + \frac{1}{2} \sum_{j=1}^{i-1} x_{ji} + \frac{1}{2} \sum_{j=i+1}^n x_{ij}$$

$$p_i = P_{ii} + \frac{1}{2} \sum_{j=i+1}^n H_{ij}$$

Example for n=3 alleles:

$$i = 1 \rightarrow p_i = x_{11} + \cancel{\frac{1}{2} \sum_{j=1}^0 x_{j1}} + \frac{1}{2} \sum_{j=2}^3 x_{1j} \rightarrow p_1 = x_{11} + \frac{1}{2}x_{12} + \frac{1}{2}x_{13}$$

$$i = 2 \rightarrow p_2 = x_{22} + \frac{1}{2} \sum_{j=1}^1 x_{j2} + \frac{1}{2} \sum_{j=3}^3 x_{ij} \rightarrow p_2 = x_{22} + \frac{1}{2}x_{12} + \frac{1}{2}x_{23}$$

$$i = 3 \rightarrow p_3 = x_{33} + \frac{1}{2} \sum_{j=1}^2 x_{j3} + \cancel{\frac{1}{2} \sum_{j=4}^3 x_{3j}} \rightarrow p_3 = x_{33} + \frac{1}{2}x_{13} + \frac{1}{2}x_{23}$$

Example for n=4 alleles:

$$i = 1 \rightarrow p_i = x_{11} + \cancel{\frac{1}{2} \sum_{j=1}^0 x_{j1}} + \frac{1}{2} \sum_{j=2}^4 x_{1j} \rightarrow p_1 = x_{11} + \frac{1}{2}x_{12} + \frac{1}{2}x_{13} + \frac{1}{2}x_{14}$$

$$i = 2 \rightarrow p_2 = x_{22} + \frac{1}{2} \sum_{j=1}^1 x_{j2} + \frac{1}{2} \sum_{j=3}^4 x_{ij} \rightarrow p_2 = x_{22} + \frac{1}{2}x_{12} + \frac{1}{2}x_{23} + \frac{1}{2}x_{24}$$

$$i = 3 \rightarrow p_3 = x_{33} + \frac{1}{2} \sum_{j=1}^2 x_{j3} + \frac{1}{2} \sum_{j=4}^4 x_{3j} \rightarrow p_3 = x_{33} + \frac{1}{2}x_{13} + \frac{1}{2}x_{23} + \frac{1}{2}x_{34}$$

$$i = 4 \rightarrow p_4 = x_{44} + \frac{1}{2} \sum_{j=1}^3 x_{j4} + \cancel{\frac{1}{2} \sum_{j=5}^4 x_{4j}} \rightarrow p_4 = x_{44} + \frac{1}{2}x_{14} + \frac{1}{2}x_{24} + \frac{1}{2}x_{34}$$

The order of the alleles does not matter and both orientations can be combined in the same heterozygote. (e.g. $H_{12} \rightarrow x_{12} + x_{21}$).

$$i = 1 \rightarrow p_1 = x_{11} + \frac{1}{2}x_{12} + \frac{1}{2}x_{13}$$

$$i = 1 \rightarrow p_1 = x_{11} + \frac{1}{2}x_{12} + \frac{1}{2}x_{21} + \frac{1}{2}x_{13} + \frac{1}{2}x_{31}$$

$$i = 2 \rightarrow p_2 = x_{22} + \frac{1}{2}x_{12} + \frac{1}{2}x_{23}$$

$$i = 2 \rightarrow p_2 = x_{22} + \frac{1}{2}x_{12} + \frac{1}{2}x_{21} + \frac{1}{2}x_{13} + \frac{1}{2}x_{31}$$

$$i = 3 \rightarrow p_3 = x_{33} + \frac{1}{2}x_{13} + \frac{1}{2}x_{23}$$

$$i = 3 \rightarrow p_3 = x_{33} + \frac{1}{2}x_{13} + \frac{1}{2}x_{31} + \frac{1}{2}x_{23} + \frac{1}{2}x_{32}$$

The total frequency of homozygotes is given by

$$G = \sum_{i=1}^k p_i^2$$

G is called the **homozygosity** of the locus.

The **heterozygosity** of the locus is given by

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

Note : The definition of heterozygosity uses only allele frequency,
not genotype frequencies!

In statistics, **degrees of freedom (df)** refer to the number of values in the final calculation of a statistic that are free to vary. They are a crucial concept when working with various statistical tests and distributions. Degrees of freedom represent the flexibility in a sample or data set, allowing for a more accurate assessment of uncertainty in statistical calculations. For example, in a chi-square test, degrees of freedom are related to the number of categories in a contingency table. Understanding degrees of freedom is essential for interpreting the results of statistical tests and making valid inferences.

$$df = n(\text{genotypes}) - n(\text{alleles})$$

PopGen ▷ Hardy-Weinberg Principle

At a locus with n alleles, how many different genotypes are there?

A_1	$A_1 A_2$	$A_1 A_2 A_3$	$A_1 A_2 A_3 A_4$
A_1A_1	A_1A_1 A_1A_2 A_2A_2	A_1A_1 A_1A_2 A_2A_2 A_1A_3 A_3A_3 A_2A_3	A_1A_1 A_1A_2 A_2A_2 A_1A_3 A_3A_3 A_1A_4 A_4A_4 A_2A_3 A_2A_4 A_3A_4
$n=1$	$n=3$	$n=6$	$n=10$

$$n(\text{genotypes}) = \frac{n(n+1)}{2}$$

Example for Alkaline phosphatase locus with 3 alleles (Harris 1966).

Genotype*	Number	Frequency
SS	141	0.4247
SF	111	0.3343
FF	28	0.0843
SI	32	0.0964
FI	15	0.0452
II	5	0.0151
Total	332	1.0000

*Alkaline phosphatase (ALP, ALKP) is a hydrolase enzyme responsible for removing phosphate groups of molecules, including nucleotides and proteins.

PopGen ▷ Hardy-Weinberg Principle

Genotype	Number	Frequency
SS	141	0.4247
SF	111	0.3343
FF	28	0.0843
SI	32	0.0964
FI	15	0.0452
II	5	0.0151
Total	332	1.0000

→ observed genotype frequencies

Alleles frequencies:

$$f(S) = 0.4247 + \frac{1}{2}0.3343 + \frac{1}{2}0.0964 = 0.6401$$

Genotype	Number	Frequency
SS	141	0.4247
SF	111	0.3343
FF	28	0.0843
SI	32	0.0964
FI	15	0.0452
II	5	0.0151
Total	332	1.0000

Alleles frequencies:

$$f(S) = 0.4247 + \frac{1}{2} 0.3343 + \frac{1}{2} 0.0964 = 0.6401$$

$$f(F) = 0.0843 + \frac{1}{2} 0.3343 + \frac{1}{2} 0.0452 = 0.2741$$

$$f(I) = 0.0151 + \frac{1}{2} 0.0964 + \frac{1}{2} 0.0452 = 0.0859$$

$$f(S) = 0.6401$$

$$f(F) = 0.2741$$

$$f(I) = 0.0859$$

Expected frequency for homozygotes:

$$f(SS)_{\text{expected}} = f(S)^2$$

$$f(FF)_{\text{expected}} = f(F)^2$$

$$f(II)_{\text{expected}} = f(I)^2$$

Expected frequency for heterozygotes:

$$f(SF)_{\text{expected}} = 2 * (f(S) * f(F))$$

$$f(SI)_{\text{expected}} = 2 * (f(S) * f(I))$$

$$f(FI)_{\text{expected}} = 2 * (f(F) * f(I))$$

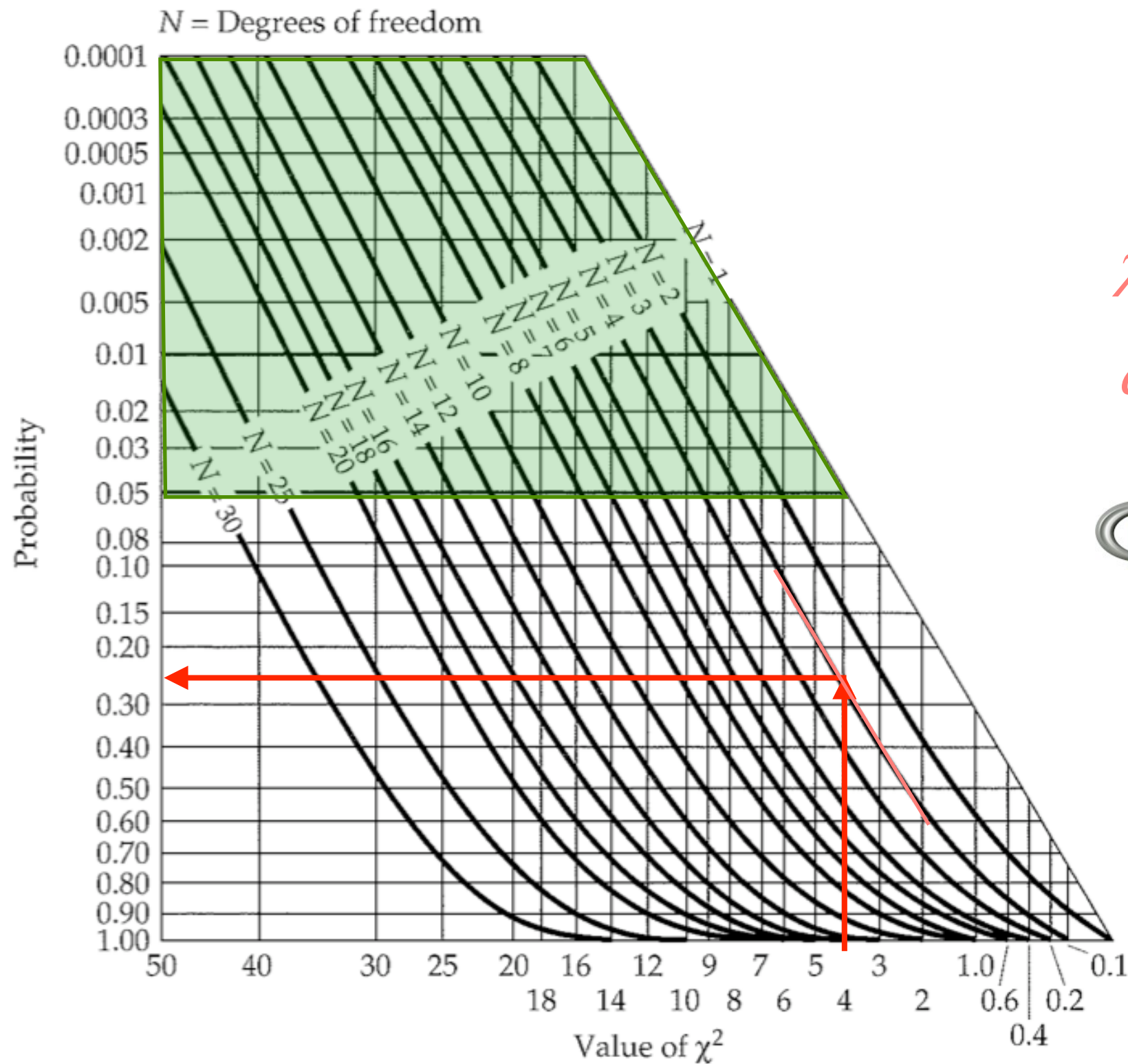
PopGen ▷ Hardy-Weinberg Principle

Genotype	Number	GT-Frequency	expected Frequency	expected Number
SS	141	0.4247	0.4097	136.03
SF	111	0.3343	0.3509	116.50
FF	28	0.0843	0.0751	24.94
SI	32	0.0964	0.1100	36.51
FI	15	0.0452	0.0471	15.63
II	5	0.0151	0.0074	2.45
Total	332	1.0000	1.0002	332.07

PopGen ▷ Hardy-Weinberg Principle

Genotype	observed Number	expected Number	Δ^2	Δ^2 / exp
SS	141	136.03	24.7	0.18
SF	111	116.50	30.3	0.26
FF	28	24.94	9.4	0.38
SI	32	36.51	20.3	0.56
FI	15	15.63	0.4	0.03
II	5	2.45	6.5	2.65
Total	332	332.06		4.05 = χ^2

PopGen ▷ Hardy-Weinberg Principle



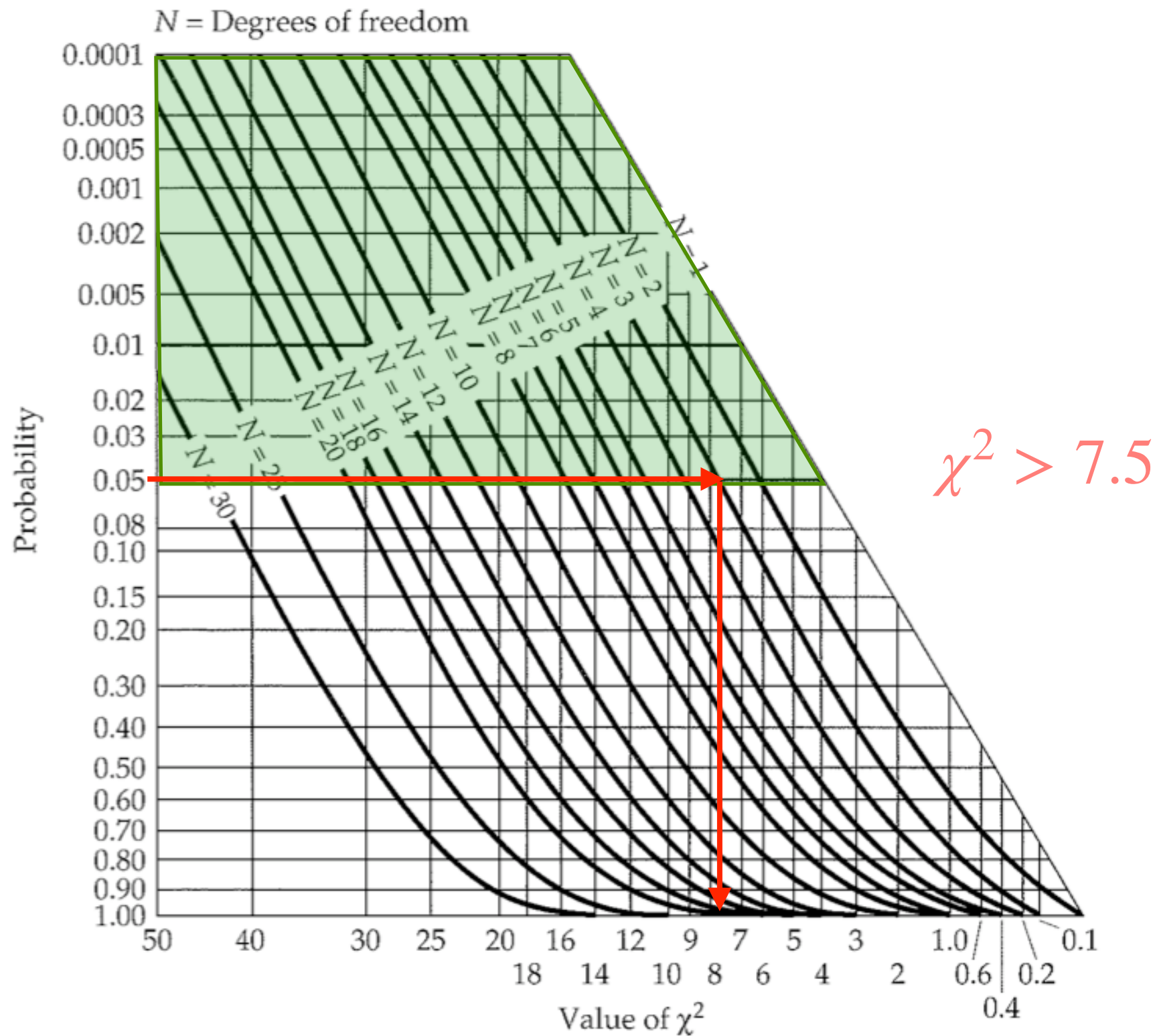
$$\chi^2 = 4.05$$

$$df = 6 - 3 = 3$$



```
1-pchisq(4.05, 3)  
[1] 0.256
```

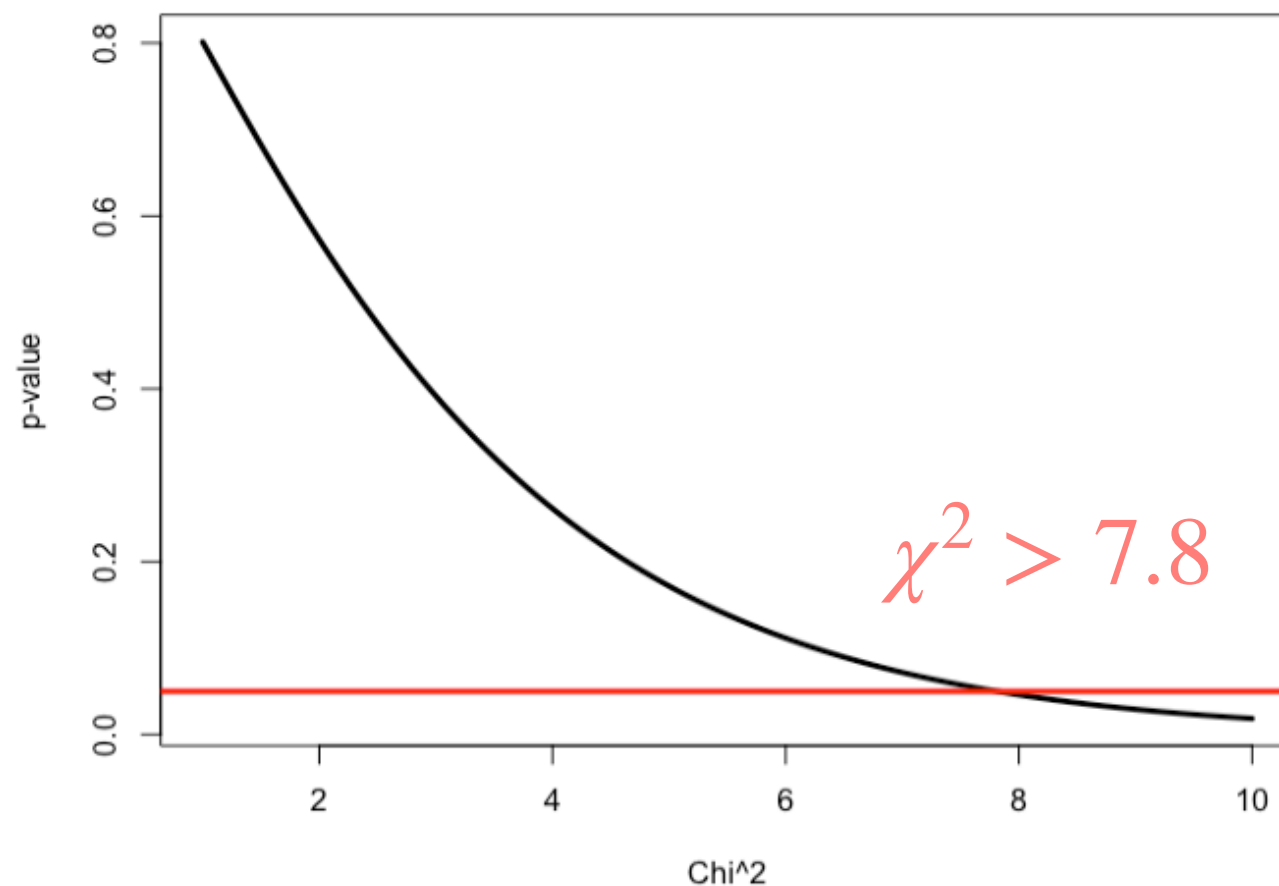
PopGen ▷ Hardy-Weinberg Principle



PopGen ▷ Hardy-Weinberg Principle



```
p.v <- function(x, df=3){  
  1-pchisq(x, df)  
}  
  
# Plot  
plot(p.v, xlim = range(1:10),  
      xlab = "Chi^2",  
      ylab = "p-value")  
abline(h = 0.05, col = "red")
```



Conclusion:

With a high probability, we assume the population is in H-W at the alkaline phosphatase locus.

The classical χ^2 test has limits!

The classical chi-squared test is commonly used to assess whether the observed genotype frequencies in a population deviate from the genotype frequencies expected under Hardy-Weinberg equilibrium (HWE). Hardy-Weinberg equilibrium is a fundamental principle in population genetics that describes the distribution of genetic variation in an idealised, non-evolving population.

In the context of testing for HWE, the limitations of the classical chi-squared test are:

- If the **sample size is too small**, the test may fail to detect subtle departures from equilibrium.
- It may **not have enough statistical power to detect differences for rare alleles**.
- When dealing with **multiple loci**, the test may not take into account interactions between loci and their combined effects on deviations from HWE.

To address some of these limitations, more advanced statistical tests and methods have been developed, such as exact tests, likelihood ratio tests and Bayesian approaches. These methods can provide more accurate assessments of departures from HWE, particularly in situations where the classical chi-squared test may fall short due to its assumptions and limitations.

The **chi-square test is only an approximation** of the actual probability distribution, and the approximation becomes poor when expected numbers are small. Therefore testing for Hardy-Weinberg proportions at loci with many alleles, such as microsatellite loci, could be a problem because many genotypes will have extremely low expected numbers.



Microsatellites Reveal Population Identity of Individual Pink Salmon to Allow Supportive Breeding of a Population at Risk of Extinction

Olsen et al. (2011)

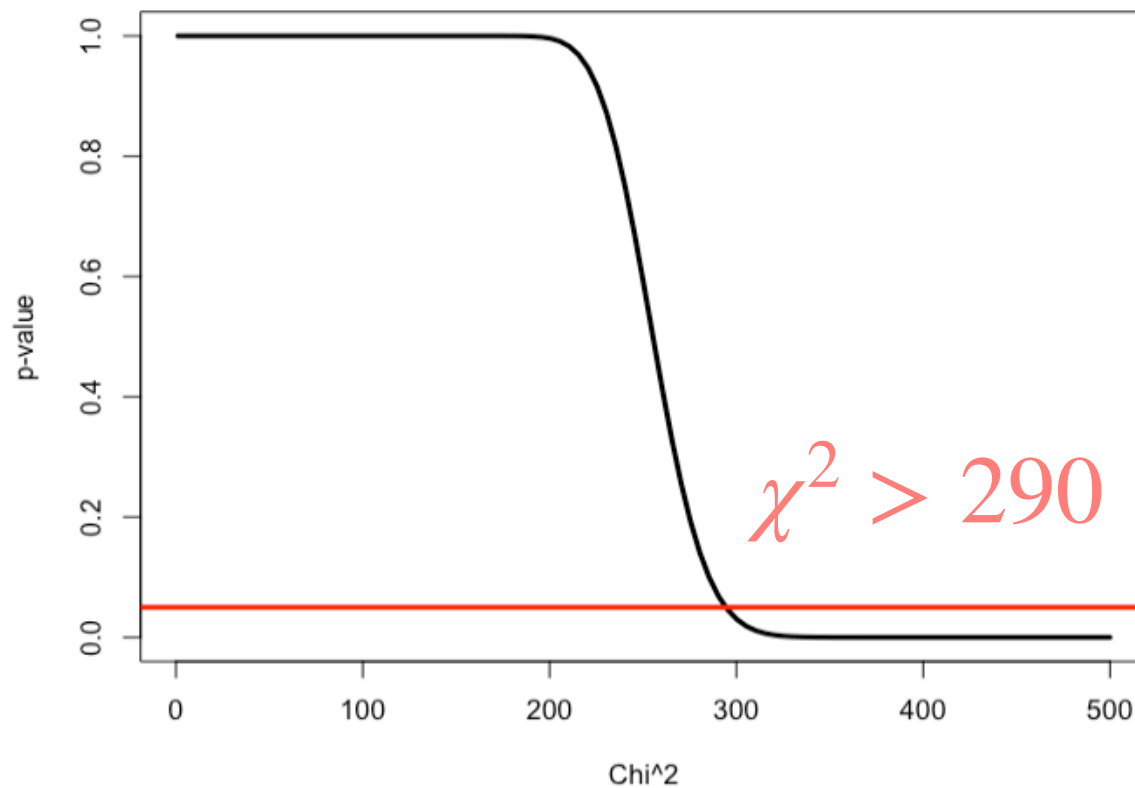
$$N_{alleles} = 23 \Rightarrow Genotypes = \frac{n(n+1)}{2} = \frac{23(23+1)}{2} = 279$$

$$d.f.: 279 - 23 = 256$$

PopGen ▷ Hardy-Weinberg Principle



```
p.v <- function(x, df=256){  
  1-pchisq(x, df)  
}  
  
# Plot  
plot(p.v, xlim = range(1:500),  
      xlab = "Chi^2",  
      ylab = "p-value")  
abline(h = 0.05, col = "red")
```



Am. J. Hum. Genet. 76:887–883, 2005

A Note on Exact Tests of Hardy-Weinberg Equilibrium

Janis E. Wigginton,¹ David J. Cutler,² and Gonçalo R. Abecasis¹

Deviations from Hardy-Weinberg equilibrium (HWE) can indicate inbreeding, population stratification and even **genotyping problems**. In samples of affected individuals, these deviations can also provide evidence of association. Tests for HWE are commonly performed using a simple χ^2 goodness-of-fit test. We show that this χ^2 test can have **inflated type I error rates (false positives)** even in relatively large samples (e.g., samples of 1,000 individuals containing ~100 copies of the minor allele). Building on previous work, we describe accurate tests for HWE, together with efficient computational methods for their implementation. Our methods adequately control for type I error in large and small samples and are computationally efficient. They have been implemented in freely available code that will be useful for assessing the quality of genotype data and for detecting genetic association or population stratification in very large datasets.

Exact test for Hardy-Weinberg equilibrium

```
x <- c(5,12,6)
names(x) <- c("AA", "AB", "BB")
HW.extest <- HWEexact(x, verbose = TRUE)
```



```
Haldane Exact test for Hardy-Weinberg equilibrium (autosomal)
using SELOME p-value
sample counts: nAA = 5 nAB = 12 nBB = 6
H0: HWE (D==0), H1: D <> 0
D = 0.2608696 p-value = 1
```

Exact Tests for Hardy–Weinberg Proportions

William R. Engels¹

Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received August 24, 2009

Accepted for publication August 29, 2009

Exact conditional tests are often required to statistically assess whether a sample of diploids comes from a population with Hardy-Weinberg proportions, or to confirm the accuracy of genotype assignments. This requirement is particularly common when the sample contains **multiple alleles and sparse data**, rendering **asymptotic methods such as the common χ^2 test unreliable**. Such an exact test can be performed using the likelihood ratio as the test statistic, rather than the more commonly used probability test. The conceptual advantages of using the likelihood **ratio are** discussed. A substantially improved algorithm is described that allows a full-enumeration exact test to be performed on sample sizes too large for previous methods. An improved Monte Carlo algorithm is also proposed for samples that preclude full enumeration. These algorithms are approximately two orders of magnitude faster than those currently in use. Finally, methods are derived to compute the number of possible samples with a given set of allele counts, a useful quantity for evaluating the feasibility of the full enumeration procedure.

Ratio test for Hardy-Weinberg equilibrium

```
x <- c(5,12,6)
names(x) <- c("AA", "AB", "BB")
HW.ratiotest <- HWLratio(x, verbose = TRUE)
```



```
Likelihood ratio test for Hardy-Weinberg equilibrium
G2 = 0.04754276 DF = 1 p-value = 0.8273956
```

PopGen ▷ Hardy-Weinberg Principle

Classical autosomal tests for Hardy-Weinberg equilibrium:

```
library("HardyWeinberg")

nGT <- c(AA = 5, AB = 12, BB = 6)

## Chi-square test
HardyWeinberg::HWChisq(nGT)

## Likelihood ratio test
HardyWeinberg::HWLratio(nGT)

## Exact test
HardyWeinberg::HWExact(nGT)

## Permutation test
HardyWeinberg::HWPerm(nGT)

## Power calculations
HardyWeinberg::HWPow(y = nGT)
```



▀ have a look at the R examples on the course website

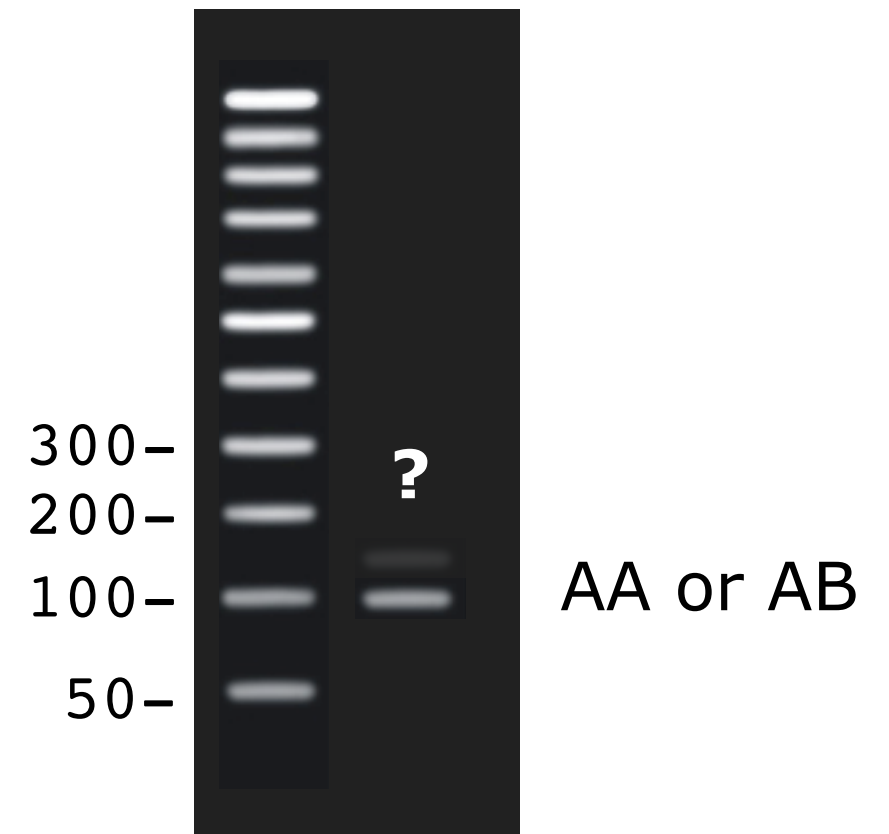
Permutation test for HWE (autosomal)

```
A3  <- c(
  SS = 141,
  SF = 111,
  FF = 28,
  SI = 32,
  FI = 15,
  II = 5
)
A3      <- toTriangular(A3)
A3.out <- HWPerm.mult(A3)
A3.out$pval
```



Null alleles at microsatellite loci result from nucleotide substitutions that prevent primers from binding (Brookfield 1996). Heterozygotes for a null allele and another allele appear as homozygotes on a gel.

The presence of null alleles results in an apparent excess of homozygotes relative to Hardy-Weinberg proportions. Brookfield (1996) discusses the estimation of null allele frequencies in the case of more than three alleles. Kalinowski and Taper (2006) have presented a maximum likelihood approach for estimating null allele frequencies at microsatellite loci.



Source: Allendorf & Luikart (2007) Conservation and the Genetics of Populations