Bioinformatics and Statistical Genetics

Elective specialization for MDS/MIRI/MAI students

Marta Castellano

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya Barcelona, Spain

marta.castellano@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Syllabus

Bioinformatics and **Statistical Genetics**

1. Introduction to statistical genetics
2. Hardy-Weinberg equilibrium 14 November 2023
3. Linkage disequilibrium and haplotype estimation
4. Population substructure
5. Genetic association analysis
6. Relatedness analysis (allele sharing)

Tuesdays 5-8pm

Content

Hardy-Weinberg Equilibrium (HWE)

1. Introduction
2. Graphics & tests
3. Disequilibrium measures
4. Multiple alleles
5. Other cases
    5.1. X chromosome
    5.2. HWE genome-wide
6. Computer exercise

Hardy-Weinberg Equilibrium (HWE)

Explain Like I'm 5

The Hardy-Weinberg Equilibrium is fundamental in **population** genetics:

When a population is in Hardy-Weinberg equilibrium for a gene, it is not evolving, and allele frequencies will stay the same across generations.

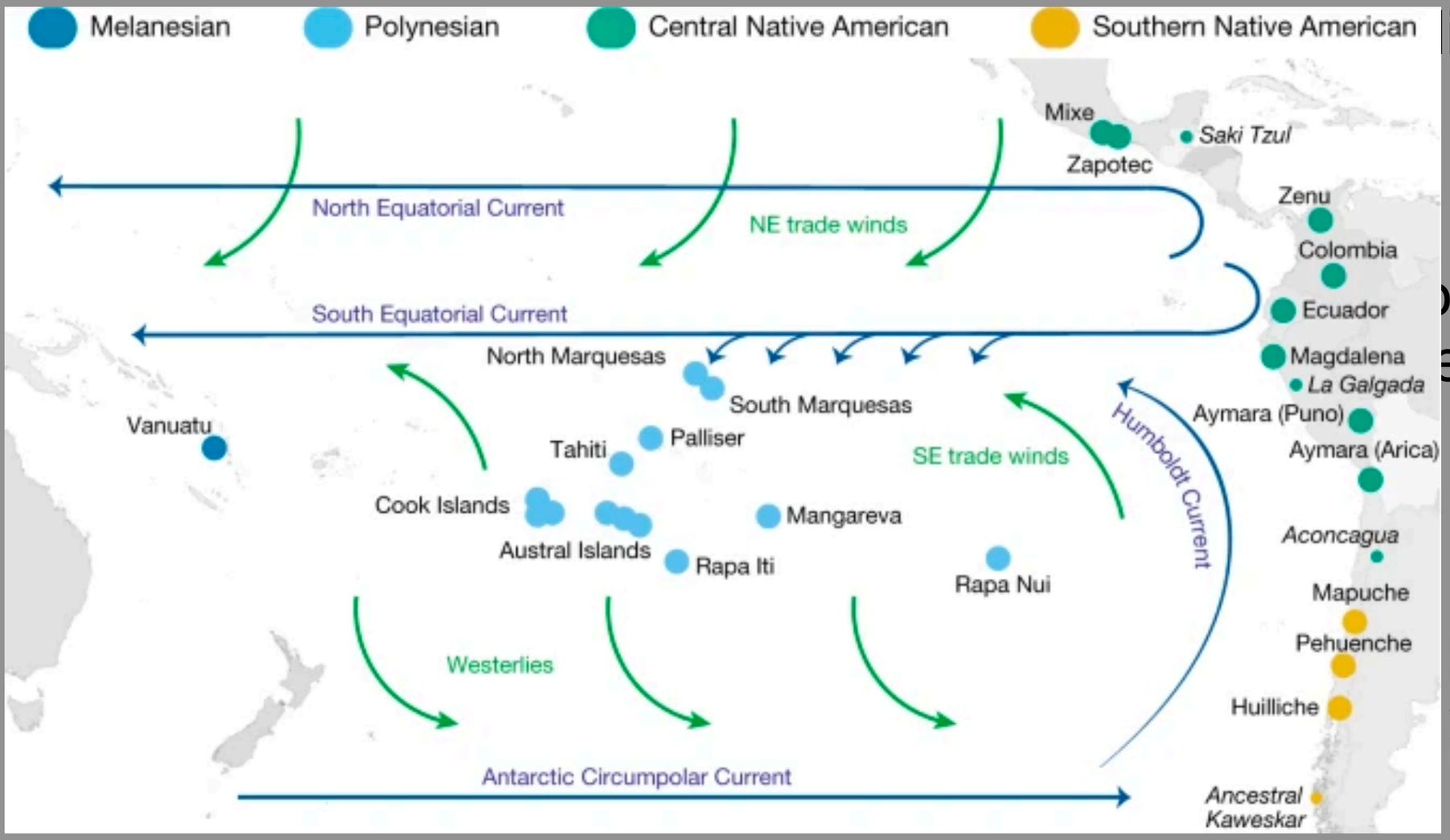
Thus, the amount of genetic variation in a population will remain constant from one generation to the next in the absence of **disturbing factors**.

Hardy-Weinberg Equilibrium (HWE)

Explain Like I'm 5

population is often defined as a set of organisms in which any pair of members can breed together….living in a defined geographic region at a specific point in time.

The Hardy-Weinberg Equilibrium is fundamental in **population** genetics:



ibrium for a
will stay the

ulation will
ext in the absence

Ioannidis, A.G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Miquel-Poblete, J.F., Moreno-Mayar, J.V., Rodríguez-Rodríguez, J.E., Quinto-Cortés, C.D., Auckland, K., Parks, T. and Robson, K., 2020. **Native American gene flow into Polynesia predating Easter Island settlement.** Nature, 583(7817), pp.572-577.

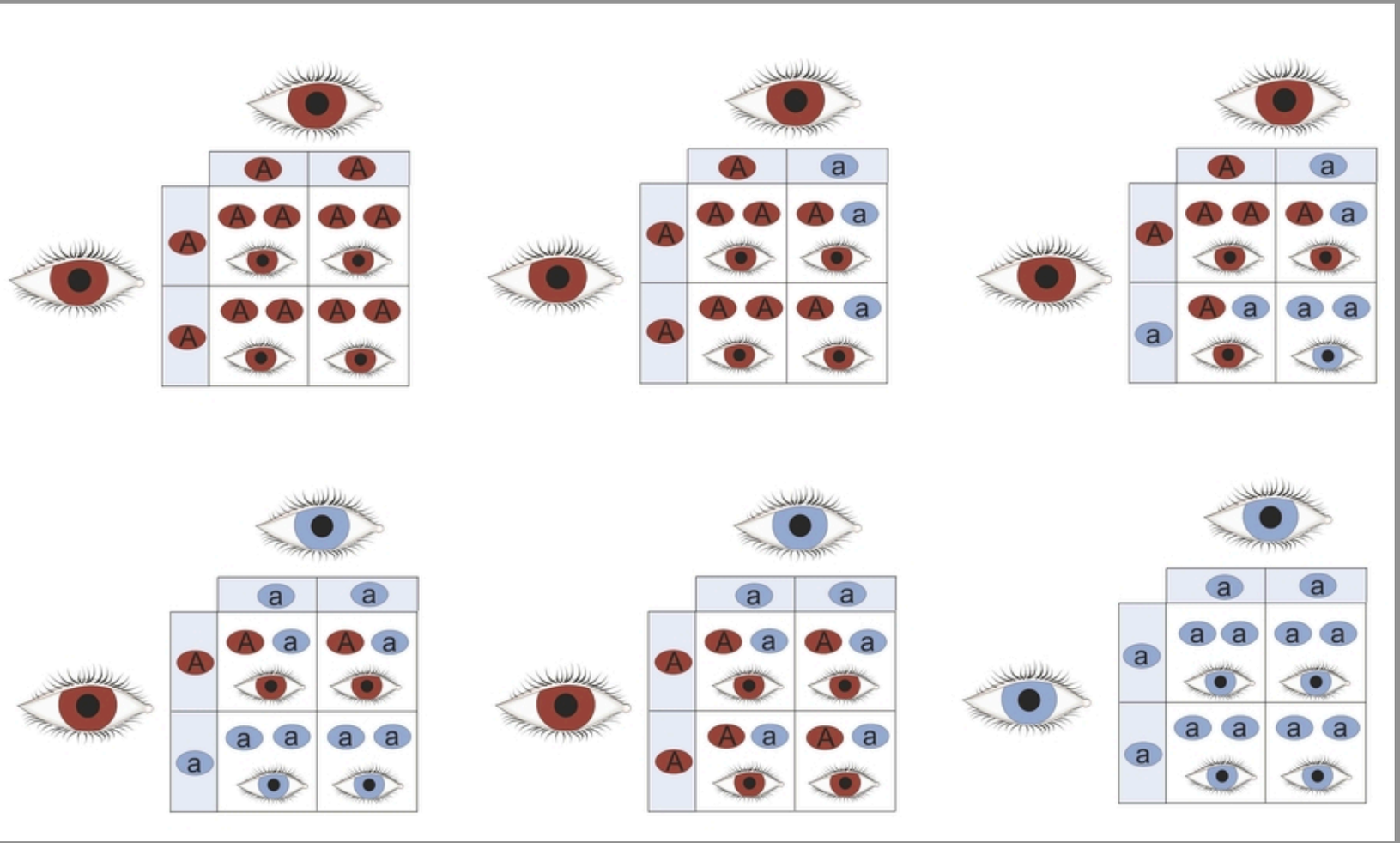
Hardy-Weinberg Equilibrium (HWE)

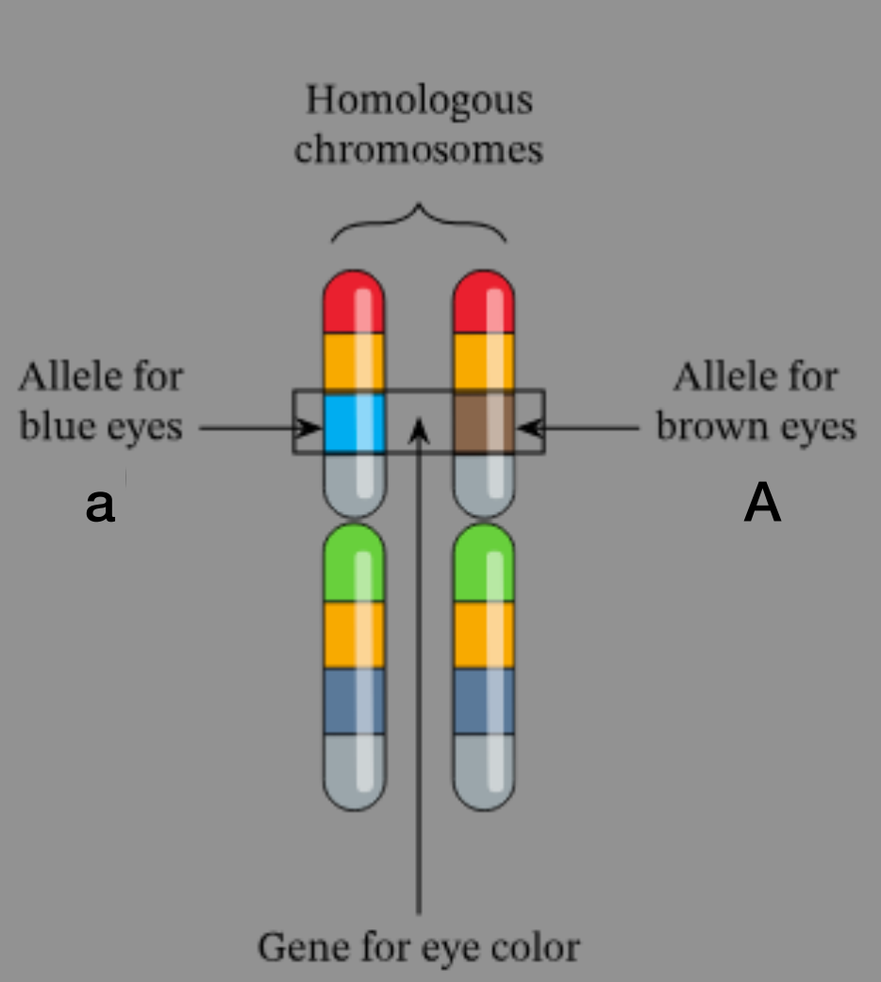
Why? (Still ELI5)

Recall that alleles can be **dominant**, **recessive** or **codominant**…and consider a bi-allelic genetic variant.

Will the dominant allele increase its frequency over time within a population?

Example: inheritance of eye color with Punnett squares





Hardy-Weinberg Equilibrium (HWE)

Why?

G.H. Hacđi (1877-1947) & W. Venbơc (1862 - 1937)

- Hardy, G.H. (1908) Mendelian proportions in a mixed population. Science 28: 49-50.
- Weinberg, W. (1908) Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg. 64:369-382.

”In a word, there is not the slightest foundation for the idea that a dominant character should show a tendency to spread over the whole population, or that a recessive should tend to die out.”

”Thus we obtain under the influence of panmixis (uniform random fertilization) in each generation the same proportion of pure and hybrid types ...”

School of Economics and Political Science, to which he was appointed in 1903, retains the readership in geography, to which, under its then title, he was appointed in 1902.

DISCUSSION AND CORRESPONDENCE

MENDELIAN PROPORTIONS IN A MIXED POPULATION

TO THE EDITOR OF SCIENCE: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making.

In the *Proceedings of the Royal Society of Medicine* (Vol. I., p. 165) Mr. Yule is reported to have suggested, as a criticism of the Mendelian position, that if brachydactyly is dominant "in the course of time one would expect, in the absence of counteracting factors, to get three brachydactylous persons to one normal."

It is not difficult to prove, however, that such an expectation would be quite groundless. Suppose that Aa is a pair of Mendelian characters, A being dominant, and that in any given generation the numbers of pure dominants (AA), heterozygotes (Aa), and pure recessives (aa) are as $p:2q:r$. Finally, suppose that the numbers are fairly large, so that the mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as

$$(p+q)^2 : 2(p+q)(q+r) : (q+r)^2,$$

or as $p_1 : 2q_1 : r_1$, say.

The interesting question is—in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is $q^2 = pr$. And since $q_1^2 = p_1 r_1$, whatever the values of p, q and r may be, the distribution will in any case continue unchanged after the second generation.

Suppose, to take a definite instance, that A is brachydactyly, and that we start from a population of pure brachydactylous and pure normal persons, say in the ratio of 1:10,000. Then $p=1$, $q=0$, $r=10{,}000$ and $p_1=1$, $q_1=10{,}000$, $r_1=100{,}000{,}000$. If brachydactyly is dominant, the proportion of brachydactylous persons in the second generation is 20,001:100,020,001, or practically 2:10,000, twice that in the first generation; and this proportion will afterwards have no tendency whatever to increase. If, on the other hand, brachydactyly were recessive, the proportion in the second generation would be 1:100,020,001, or practically 1:100,000,000, and this proportion would afterwards have no tendency to decrease.

In a word, there is not the slightest foundation for the idea that a dominant character should show a tendency to spread over a whole population, or that a recessive should tend to die out.

I ought perhaps to add a few words on the effect of the small deviations from the theoretical proportions which will, of course, occur in every generation. Such a distribution as $p_1 : 2q_1 : r_1$, which satisfies the condition $q_1^2 = p_1 r_1$, we may call a *stable* distribution. In actual fact we shall obtain in the second generation not $p_1 : 2q_1 : r_1$ but a slightly different distribution $p_1' : 2q_1' : r_1'$, which is not "stable." This should, according to theory, give us in the third generation a "stable" distribution $p_2 : 2q_2 : r_2$, also differing slightly from $p_1 : 2q_1 : r_1$; and so on. The sense in which the distribution $p_1 : 2q_1 : r_1$ is "stable" is this, that if we allow for the effect of casual deviations in any subsequent generation, we should, according to theory, obtain at the next generation a new "stable" distribution differing but slightly from the original distribution.

I have, of course, considered only the very simplest hypotheses possible. Hypotheses other that that of purely random mating will give different results, and, of course, if, as appears to be the case sometimes, the character is not independent of that of sex, or

Hardy-Weinberg Equilibrium (HWE)

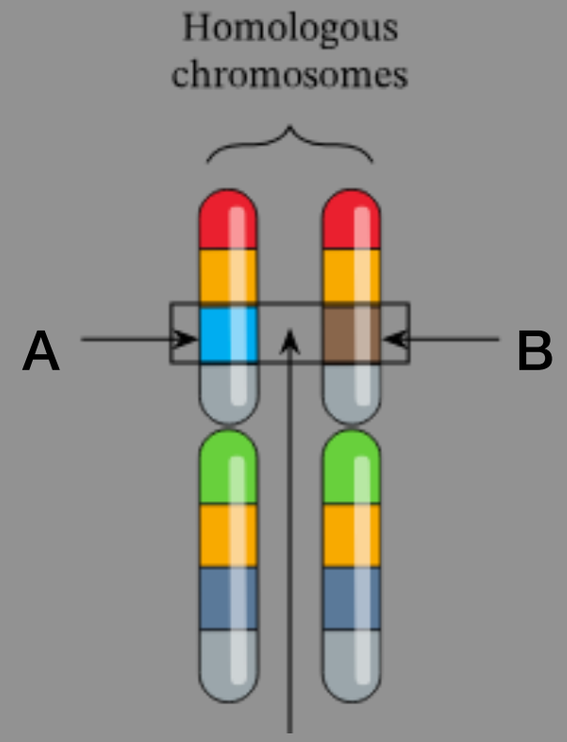
Introduction

- Consider a population of *n* individuals. With a bi-allelic genetic marker with alleles A and B and genotypes AA, AB and BB.
- The alleles A and B have frequencies p and q, so that $f_0(A) = p$ and $f_0(B) = q$.
  - Take into account that genotype frequencies sum one, so that $p + q = 1$
- The next generation an be shown in a Punnett square, with three genotypes AA, AB, BB frequencies f_{AA}, f_{AB} and f_{BB}, so that
  - $f_1(AA) = p^2 = f_0(A)^2$
  - $f_1(AB) = pq + qp = 2pq = 2f_0(A)f_0(B)$
  - $f_1(BB) = q^2 = f_0(B)^2$

		Females	
		A (p)	B (q)
Males	A (p)	AA (p^2)	AB (pq)
	B (q)	AB (qp)	BB (q^2)

- And now, we take into account that the genotype frequencies must sum one …

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



Locus of interest

Hardy-Weinberg Equilibrium (HWE)

Introduction

When a population is in Hardy-Weinberg equilibrium for a gene, it is not evolving, and allele frequencies will stay the same across generations

- In the next generation the allele frequencies read as follows
  - $f_1(A) = \frac{2f_1(AA) + f_1(AB)}{2} = f_1(AA) + \frac{1}{2}f_1(AB) = p^2 + pq = p(p+q) = p$
  - $f_1(B) = \frac{2f_1(BB) + f_1(AB)}{2} = f_1(BB) + \frac{1}{2}f_1(AB) = q^2 + pq = q(p+q) = q$

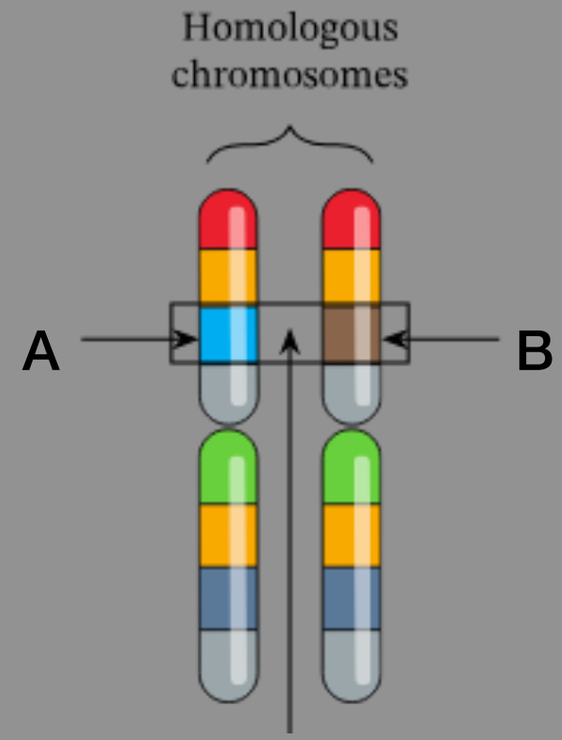
- Thus, in a single generation, equilibrium is achieved

frequency of homozygous dominant genotype

frequency of heterozygous recessive genotype

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

frequency of heterozygous genotype



Locus of interest

Hardy-Weinberg Equilibrium (HWE)

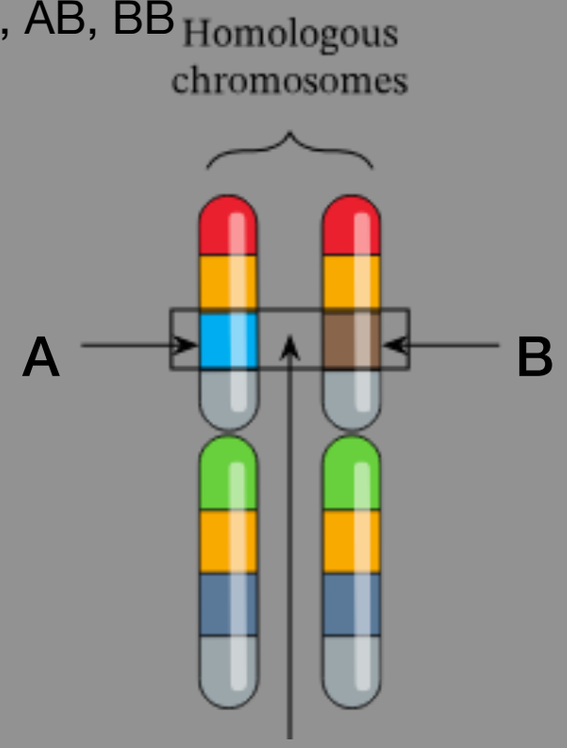
Introduction - numerical example

- Consider a population of $n = 500$ individuals. With a bi-allelic genetic marker with alleles A and B and genotypes AA, AB and BB with counts $n_{AA} = 245$, $n_{AB} = 210$ and $n_{BB} = 45$
- The alleles A and B have frequencies p and q, so that
  - $p = \frac{2 \cdot n_{AA} + n_{AB}}{2 \cdot n} = \frac{2 \cdot 245 + 210}{2 \cdot 500} = 0.7$
  - $q = \frac{2 \cdot n_{BB} + n_{AB}}{2 \cdot n} = \frac{2 \cdot 45 + 210}{2 \cdot 500} = 0.3$
- The next generation can be shown in a Punnett square, with three genotypes AA, AB, BB frequencies f_{AA}, f_{AB} and f_{BB}, so that
  - $f(AA) = p^2 = 0.49$
  - $f(AB) = pq + qp = 2pq = 2 \cdot 0.7 \cdot 0.3 = 0.42$
  - $f(BB) = q^2 = 0.3^2 = 0.09$

		Females	
		A (p)	B (q)
Males	A (p)	AA (p^2)	AB (pq)
	B (q)	AB (qp)	BB (q^2)

- And now, the genotype frequencies must sum one …

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



Locus of interest

Hardy-Weinberg Equilibrium (HWE)

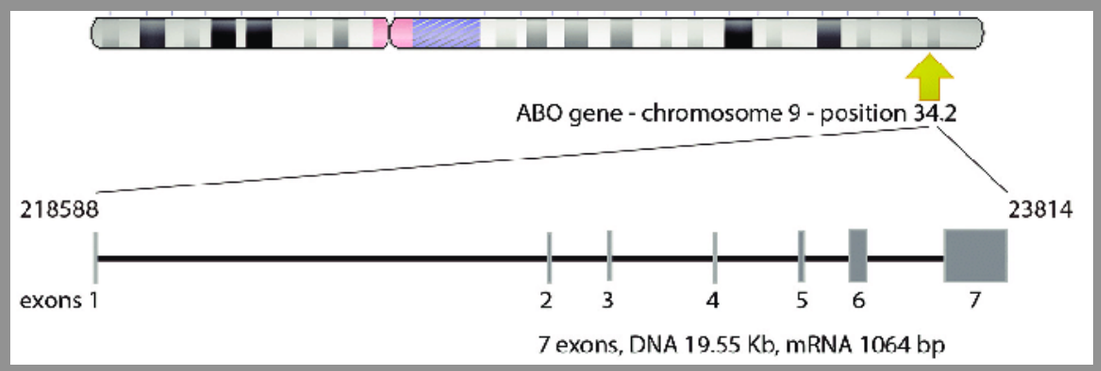
A derivation for 3 alleles

SIDE NOTE: An individual can only have 2 alleles for a given gene BUT….at the population level, human blood-group system has 3 alleles, leading to 6 possible genotypes. How many phenotypes?

- If a marker has three alleles (e.g. the blood-group system A, B and O), with frequencies p_1, p_2 and p_3.
  - Allele frequencies must sum one, so that $p_1 + p_2 + p_3 = 1$
- In the next generation…under random mating we would obtain the possible genotype frequencies:
  - $f(AA) = p_1^2$
  - $f(BB) = p_2^2$
  - $f(00) = p_3^2$ and
  - $f(AB) = 2p_1p_2$
  - $f(A0) = 2p_1p_3$
  - $f(B0) = 2p_2p_3$
- If the population is in HWE for this locus

$$p_1^2 + 2p_1p_2 + p_2^2 + 2p_2p_3 + p_3^2 + 2p_1p_3 = 1$$

		p_1 A	p_2 B	p_3 O
p_1	A	p_1^2	p_1p_2	p_1p_3
p_2	B	p_2p_1	p_2^2	p_2p_3
p_3	O	p_3p_1	p_3p_2	p_3^2

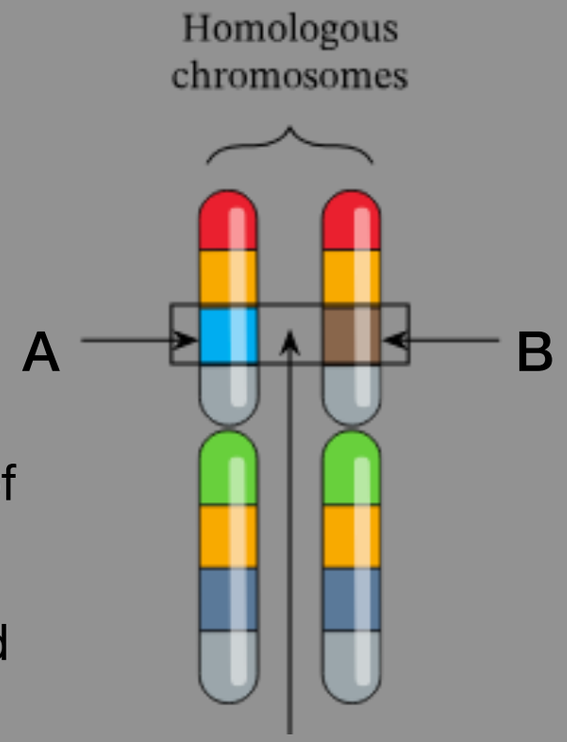


Hardy-Weinberg Equilibrium (HWE)

A derivation for multiple alleles

- Consider that given $p + q = 1$, the binomial expansion of $(p + q)^2 = 1^2$ yields to the $p^2 + 2pq + q^2 = 1$ relationship.
- Consider that given $p + q + r = 1$, the binomial expansion of $(p + q + r)^2 = 1^2$ yields to the $p_1^2 + 2p_1p_2 + p_2^2 + 2p_2p_3 + p_3^2 + 2p_1p_3 = 1$ relationship.
- If we consider $A_1 \ldots A_k$ alleles with frequencies $p_1 \ldots p_k$, the relationship of $p_1 \ldots p_k = 1$ still holds and thus, the expansion of $(p_1 \ldots p_k)^2$ would result in
  - $f(A_i, A_i) = p_i^2$ and
  - $f(A_i, A_j) = 2p_ip_j$

EXAMPLES:
- Wing shape in Drosophila



- If we consider polyploid organism with c chromosomes, two alleles A and B, frequencies p and q, then we would be working with the binomial expansion of $(p + q)^c$
- Similar generalizations can be done for more than two alleles and for polyploid systems (more than 2 chromosomes)

Hardy-Weinberg Equilibrium (HWE)

Assumptions

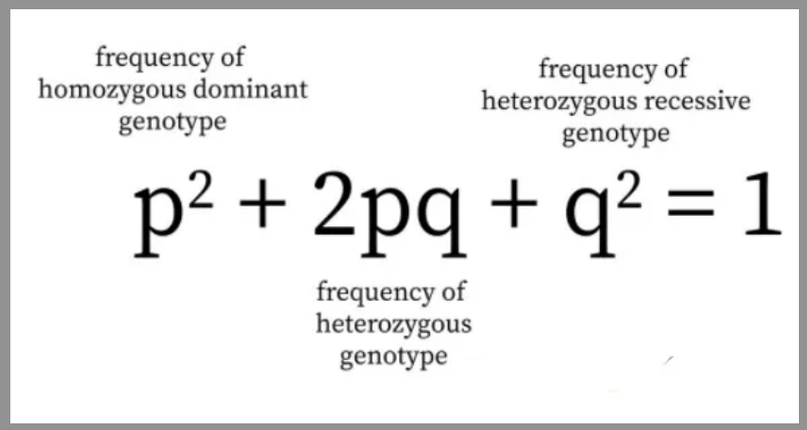
The change in frequency of an existing gene variant in the population due to random chance

- Mutation can be ignored (i.e. no genetic drift).
- There is sexual reproduction.
- There is no genotyping error.
- The organism under study is diploid.
- Allele frequencies are equal in the sexes.
- Random mating (w.r.t the trait under study).
- Population size is very (infinitely) large.
- Migration is negligible.
- Natural selection does not affect the trait under study.
- Non-overlapping generations.

Recall from ELI5:

When a population is in HWE for a gene, it is not evolving, and allele frequencies will stay the same across generations.

Thus, the amount of genetic variation in a population will remain constant from one generation to the next in the absence of **disturbing factors**.



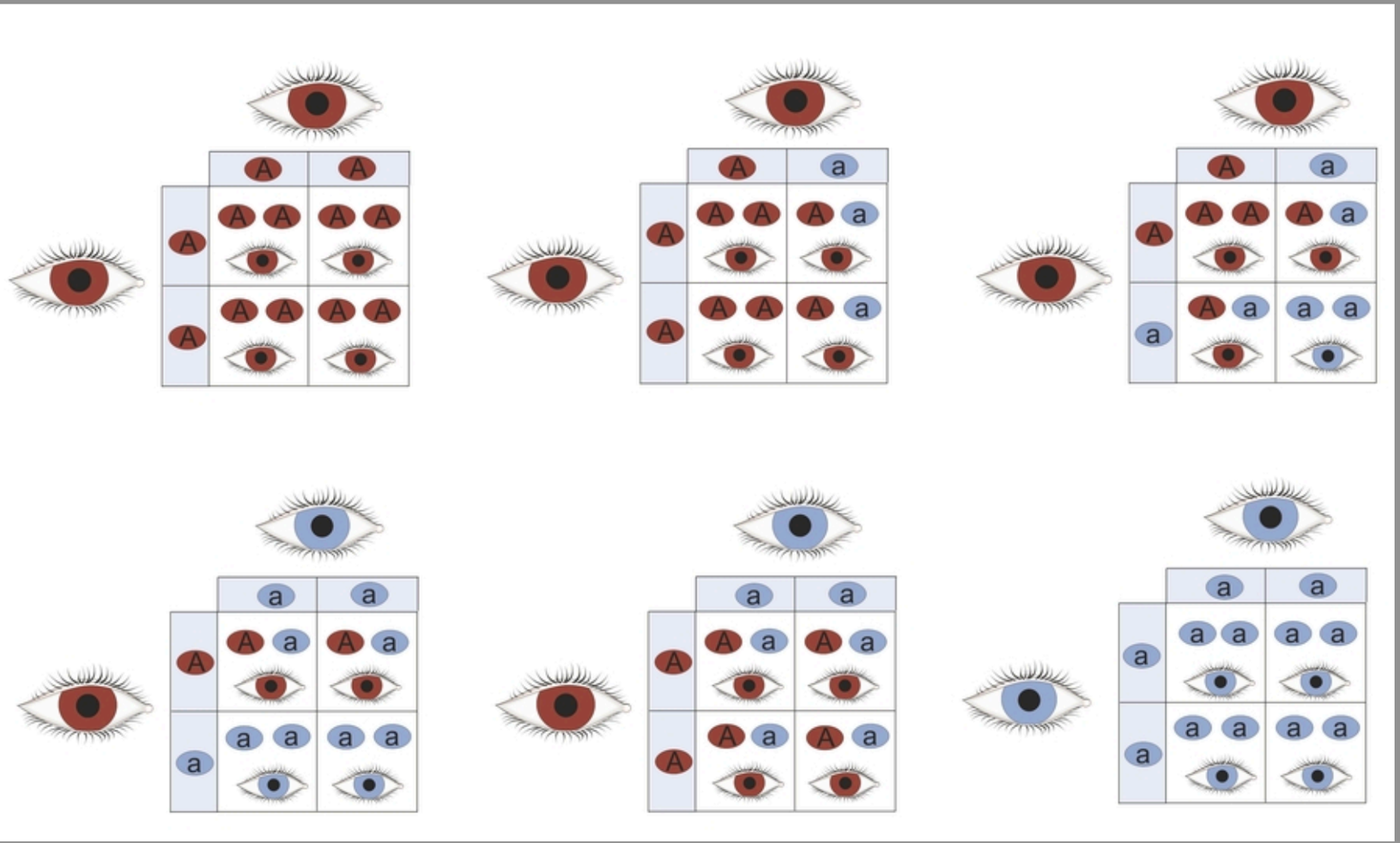
Hardy-Weinberg Equilibrium (HWE)

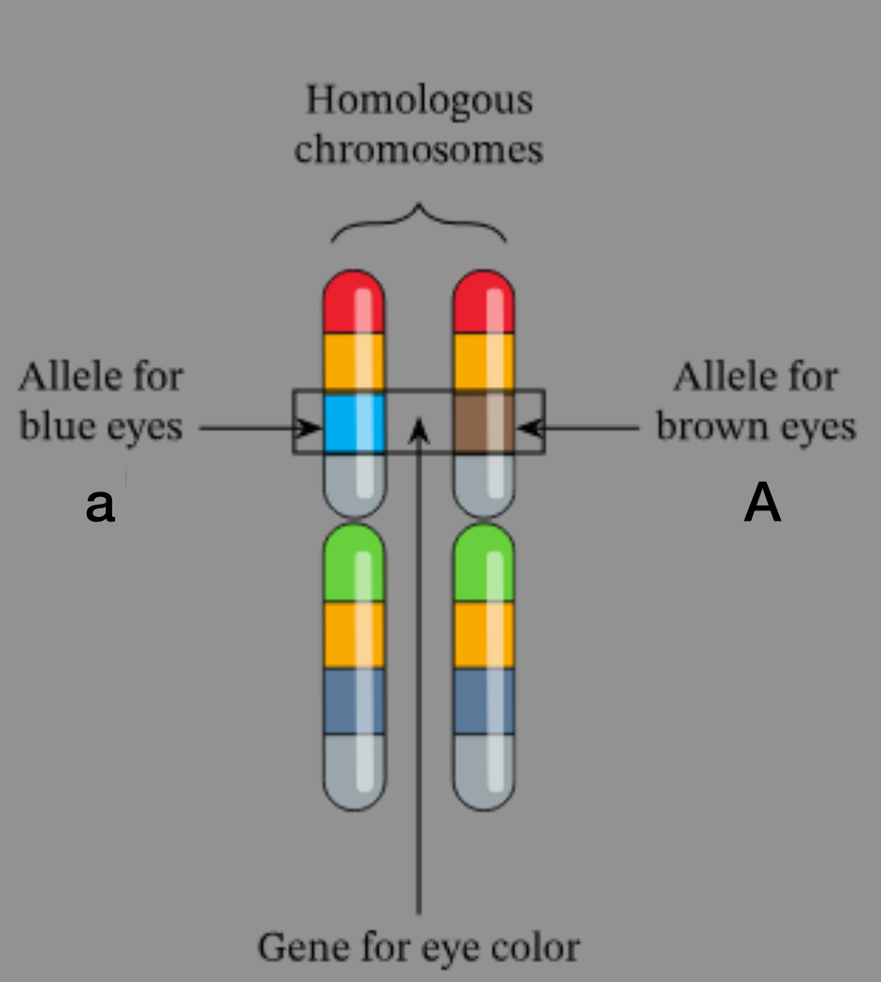
Why? (Still ELI5)

Recall that alleles can be **dominant**, **recessive** or **codominant**…and consider a bi-allelic genetic variant.

Will the dominant allele increase its frequency over time within a population?

Example: inheritance of eye color with Punnett squares





Hardy-Weinberg Equilibrium (HWE)

Why? (Still ELI5)

Recall that alleles can be **dominant**, **recessive** or **codominant**…and consider a bi-allelic genetic variant.

Will the dominant allele increase its frequency over time within a population?

- Genetic markers are, in general, expected to follow the HW law as we want them to retain predictable levels of genetic variation in the absence of forces that change allele frequencies.

- If they do not follow the law, one (or more) of the HWE assumptions is/are violated.

- The most likely cause for disequilibrium is genotyping error. Markers need to be checked for HWE as part of a quality control procedure.

- Deviation from HWP is expected (among cases) if the marker is related to disease.

Content

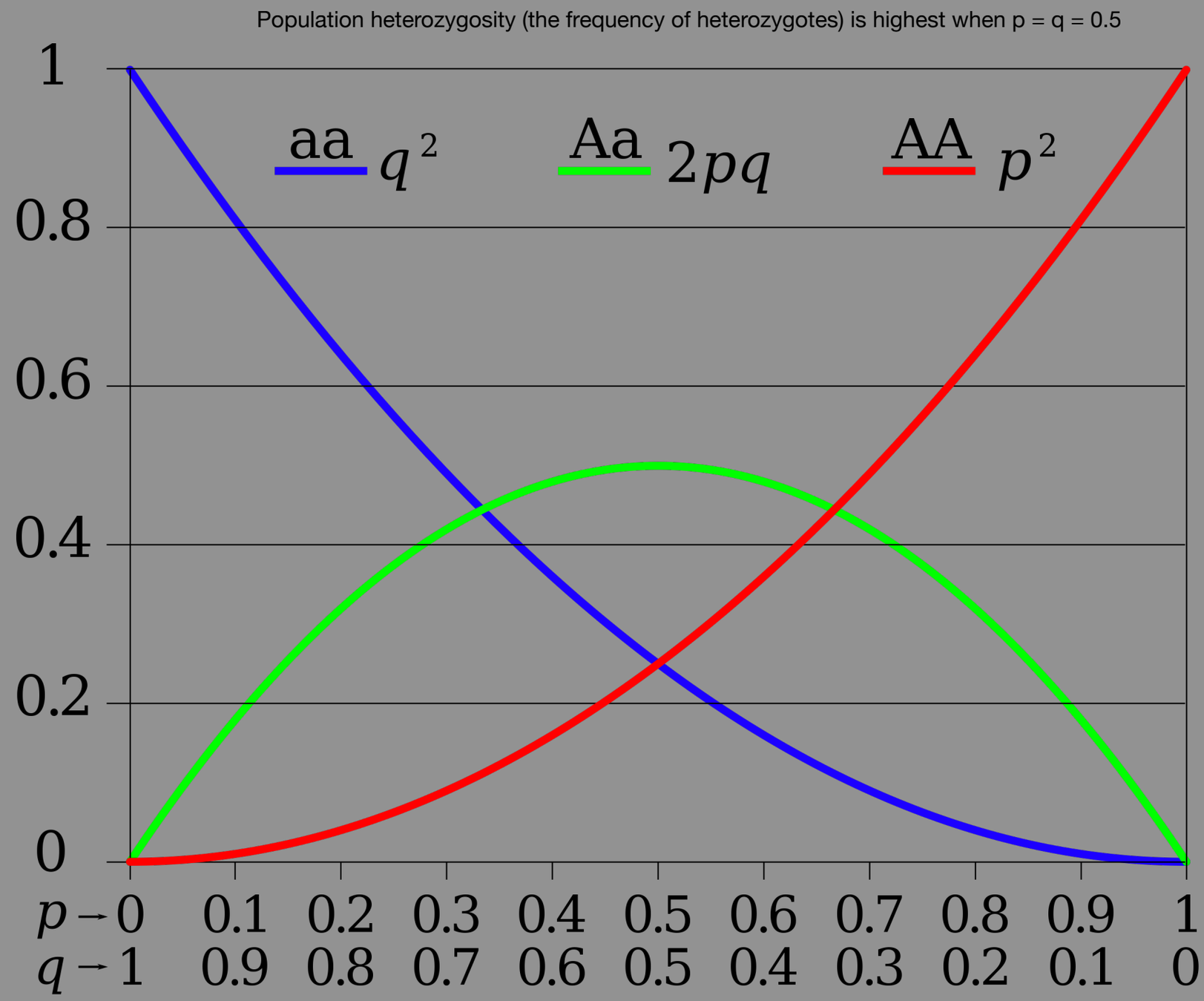
Hardy-Weinberg Equilibrium (HWE)

1. Introduction
2. Graphics & tests
3. Disequilibrium measures
4. Multiple alleles
5. Other cases
    5.1. X chromosome
    5.2. HWE genome-wide
6. Computer exercise

Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- Visualization of the Hardy–Weinberg proportions for two alleles
- The horizontal axis shows the two allele frequencies p and q and the vertical axis shows the expected genotype frequencies.
- Each line shows one of the three possible genotypes.



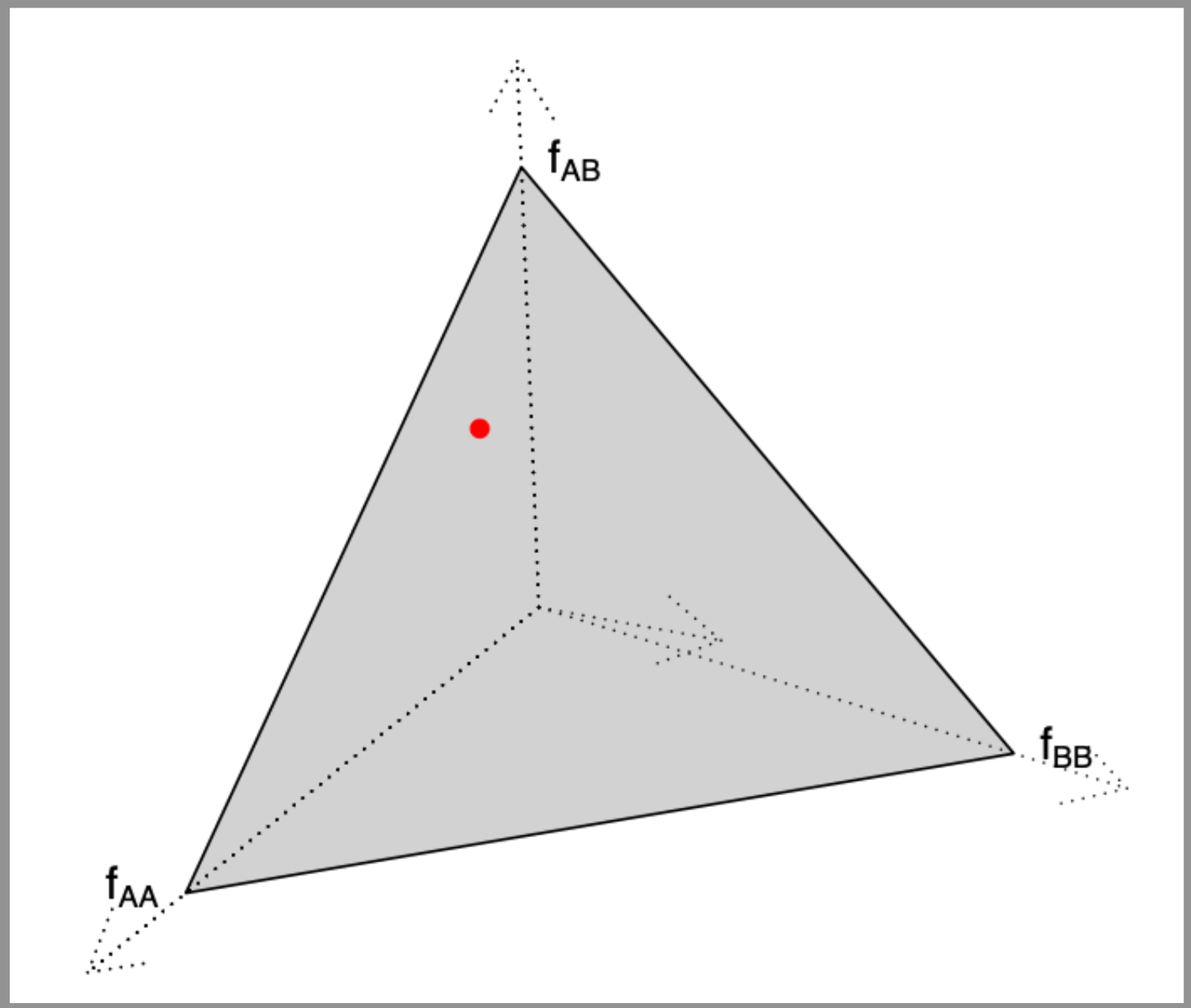
Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- Visualization of the Hardy–Weinberg proportions for two alleles

- **Ternary plot** shows the ratio of three variables that sum to a constant.

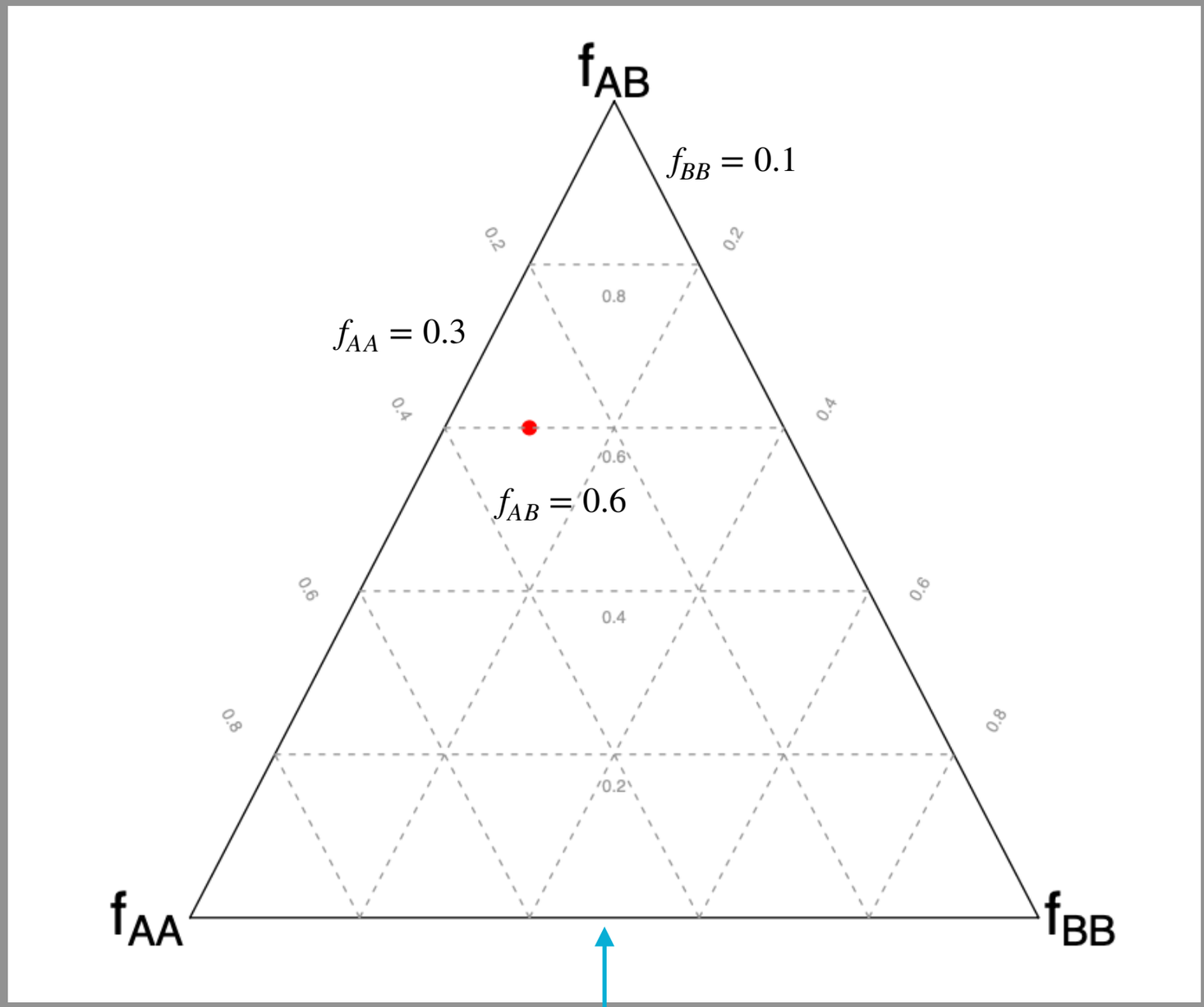
$$f_{AA} + f_{AB} + f_{BB} = 1$$



Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- How to read a ternary plot?

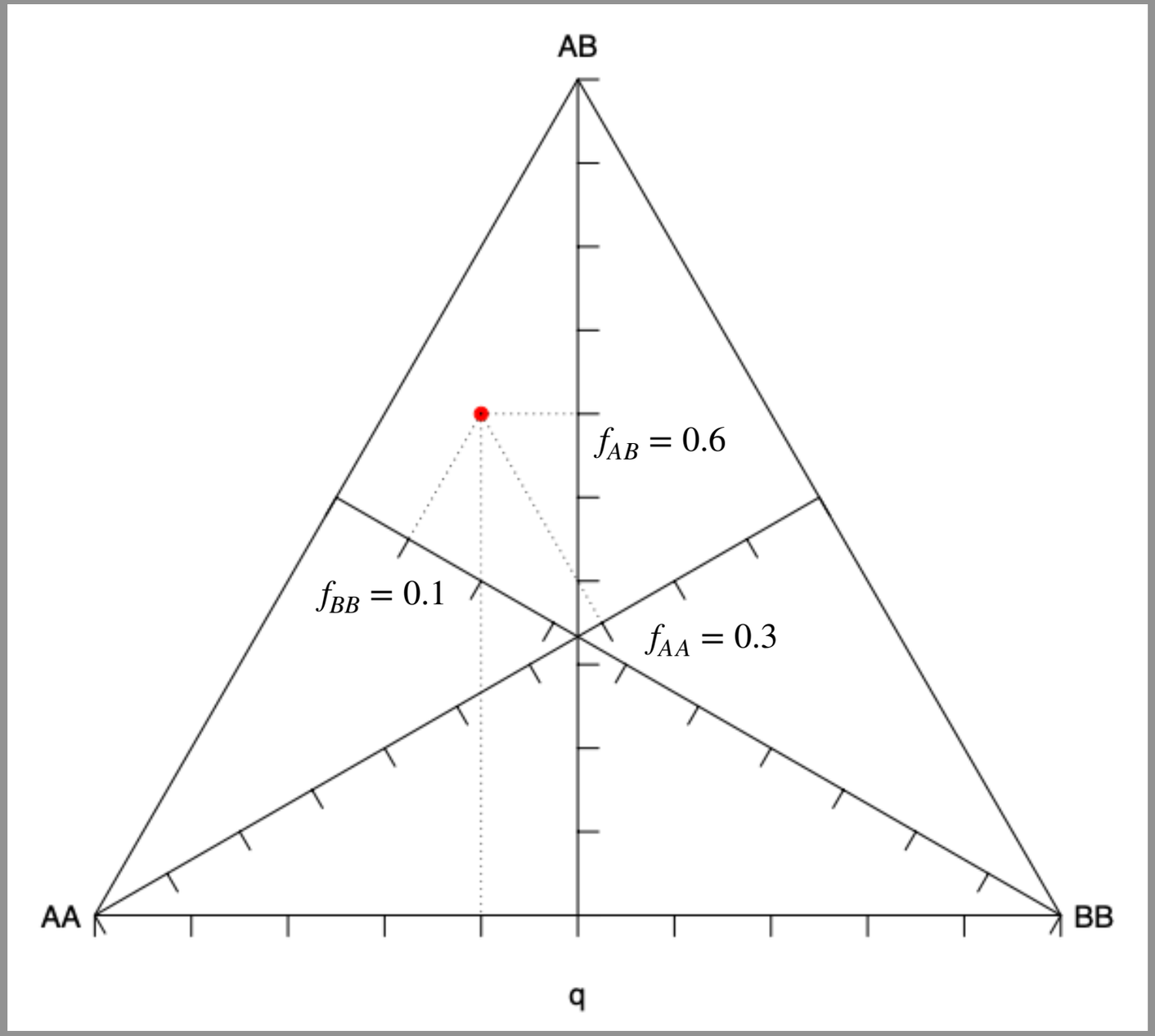


Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- How to read a ternary plot?

$$f_{AA} + f_{AB} + f_{BB} = 1$$

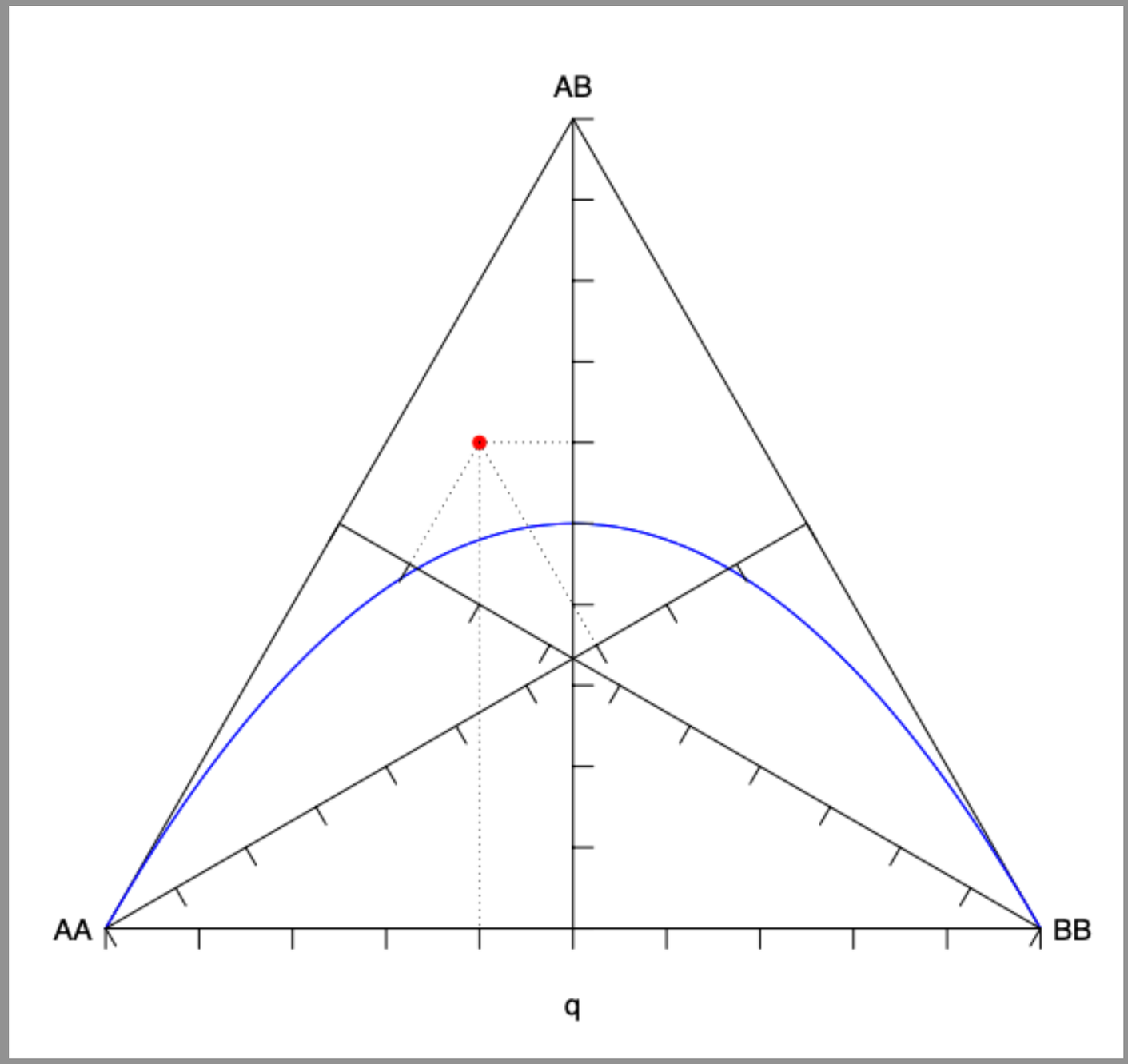


Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- How to read a ternary plot?

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

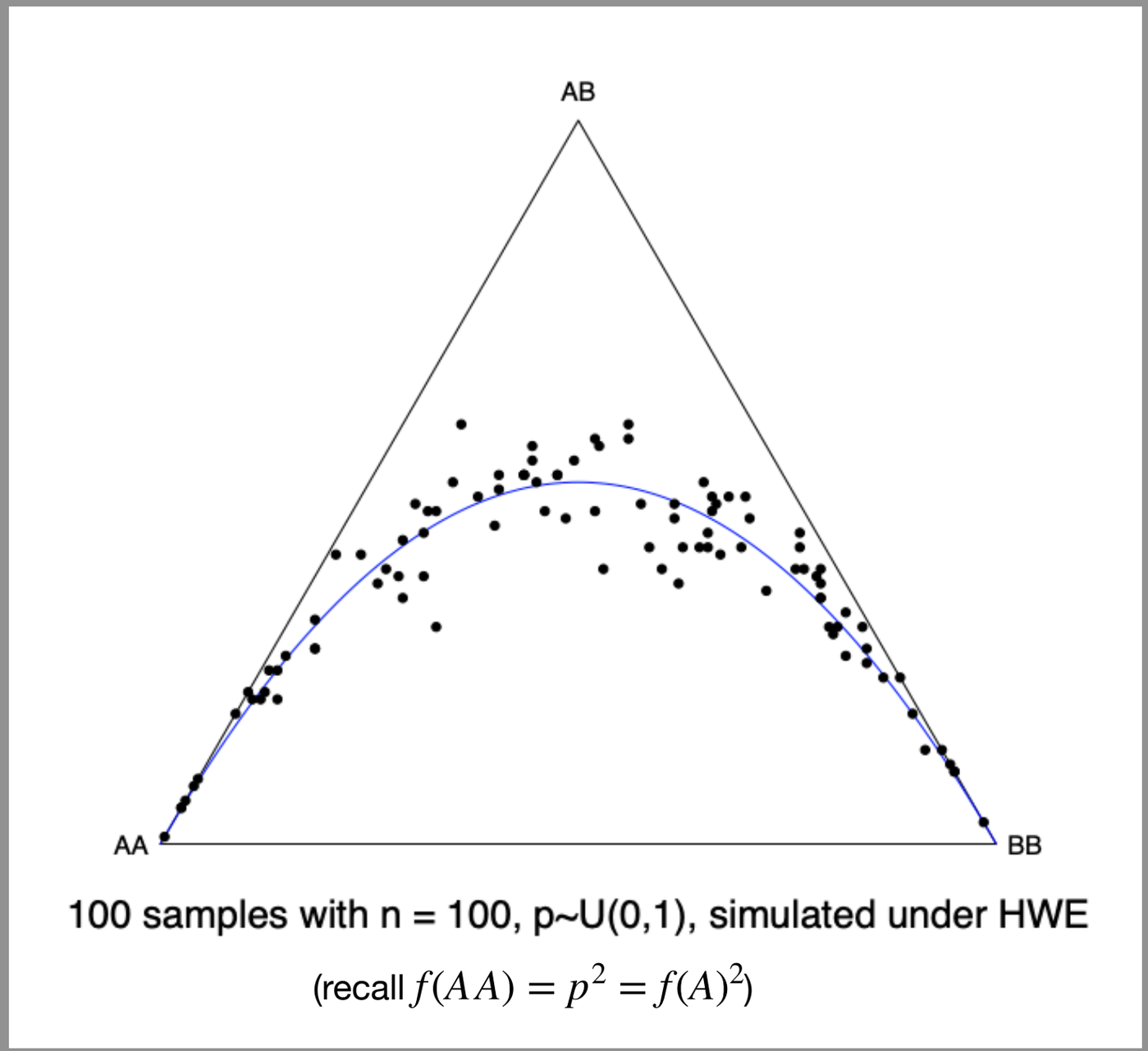


Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- How to read a ternary plot?

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



100 samples with n = 100, p~U(0,1), simulated under HWE

(recall $f(AA) = p^2 = f(A)^2$)

Content

Hardy-Weinberg Equilibrium (HWE)

1. Introduction
2. Graphics & tests
3. Disequilibrium measures
4. Multiple alleles
5. Other cases
    5.1. X chromosome
    5.2. HWE genome-wide
6. Computer exercise

Hardy-Weinberg Equilibrium (HWE)

Graphics and tests

- Testing for Equilibrium refers to the fact that once the proportions p^2, $2pq$ and q^2 are reached, allele frequencies and genotype frequencies will remain the same over the generations.

- Statistical tests for HWE test if the **hypothesis** $f(AA) = p^2, f(AB) = 2pq$, $f(BB) = q^2$ is tenable.

- Strictly speaking, statistical tests for HWE do not assess equilibrium, but test for Hardy-Weinberg proportions.

Hardy-Weinberg Equilibrium (HWE)

Statistical tests

- Pearson's χ^2 test.
- Fisher's exact test (based on $P(N_{AB} \mid N_A)$).
- Permutation test.
- …

Hardy-Weinberg Equilibrium (HWE)

Pearson's χ^2 test

- Pearson's chi-squared test is a statistical test applied to sets of **categorical data** to evaluate how likely it is that any observed difference between the sets arose by chance. Large sample sizes needed.
  - df = degree's of freedom = genotypic classes-alleles
- Let n_{AA}, n_{AB} and n_{AB} the counts for a particular set of alleles A and B (df=3-2=1).
- The expected counts under HWE

$$exp(AA) = np^2$$

$$exp(AB) = n2pq$$

$$exp(BB) = nq^2$$

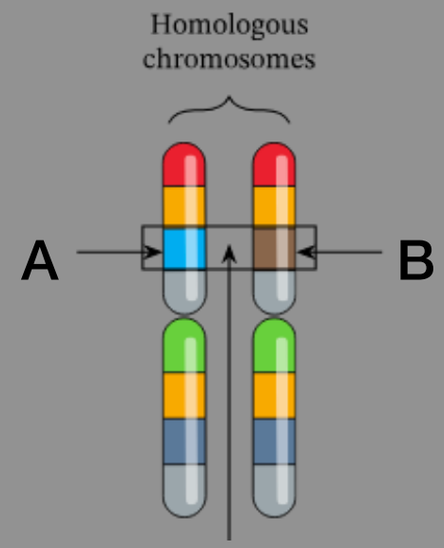
- Assumes that the test statistic follows a χ^2 distribution, and thus…a chi-square statistic for goodness of fit can be used

$$\chi^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected}$$

- If the expected counts are small, a continuity correction can be applied:

$$\chi_c^2 = \sum_{genotypes} \frac{(|observed - expected| - c)^2}{expected} \text{ for } c = 0.5$$

- And define the deviation from independence $D = \frac{1}{2}(n_{AB} - exp_{AB})$



Gene of interest

Hardy-Weinberg Equilibrium (HWE)

Pearson's χ^2 test - numerical example

- For an A/T polymorphism with $n_{AA} = 46$, $n_{AT} = 39$ and $n_{TT} = 15$ counts for a df= degrees of freedom= genotypic classes-alleles =3-2=1

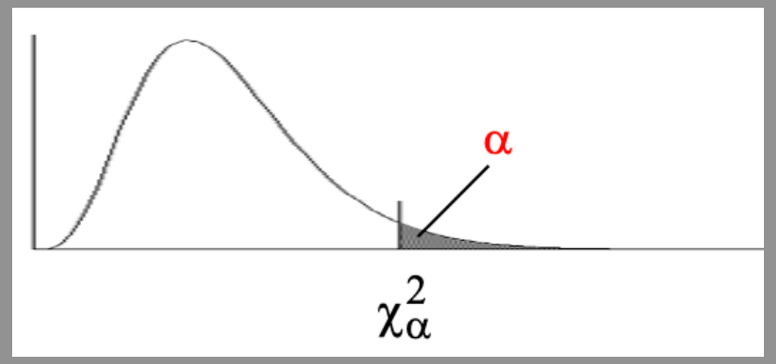
- Estimate $\hat{p}_A = \frac{n_{AA} + \frac{1}{2}n_{AT}}{n} = \frac{46 + \frac{1}{2}39}{100} = 0.655$

- The expected counts under HWE

$$exp(AA) = n\hat{p}_A^2 = 100 \cdot (0.655)^2 = 42.9025$$

$$exp(AT) = n2pq = n2\hat{p}_A(1 - \hat{p}_A) = 45.195$$

$$exp(TT) = nq^2 = n(1 - \hat{p}_A)^2 = 11.9025$$



$\chi^2 = 1.8789$ $\chi^2_\alpha = 3.841$

$\alpha = 0.05$

- A chi-square statistic for goodness of fit can be used

$$\chi^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected} =$$

$$= \frac{(46 - 42.9025)^2}{42.9025} + \frac{(39 - 45.195)^2}{45.195} + \frac{(15 - 11.9025)^2}{11.9025} = 1.8789$$

Is the variation between the observed and the expected due to chance?

If the $\chi^2$of the sample is less than $\chi^2_\alpha = 3.841$, we have failed to reject the null hypothesis… and thus, our alleles are in equilibrium

P-value = $P(\chi^2 \geq 1.8789) = 0.1704601$

Our alleles are in equilibrium in this population as we have FAILED to reject the null hypothesis

Hardy-Weinberg Equilibrium (HWE)

Pearson's χ^2 test - examples in R

- A chi-square statistic for goodness of fit can be used

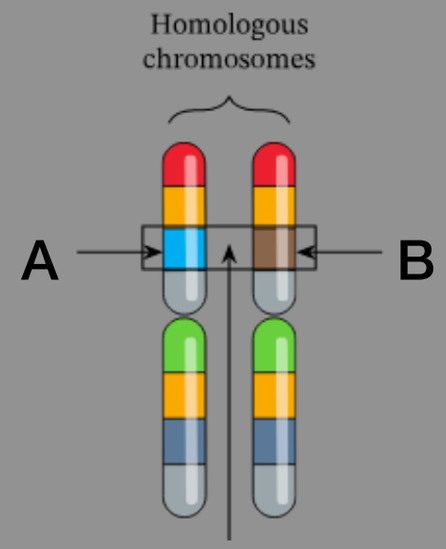
$$\chi^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected}$$

```
> library(HardyWeinberg)
> x <- c(46,39,15)
> names(x) <- c("AA","AT","TT")
> results <- HWChisq(x,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  1.878892 p-value =  0.1704601
```

- If the expected counts are small, a continuity correction can be applied:

$$\chi_c^2 = \sum_{genotypes} \frac{(|observed - expected| - c)^2}{expected} \text{ for } c = 0.5$$

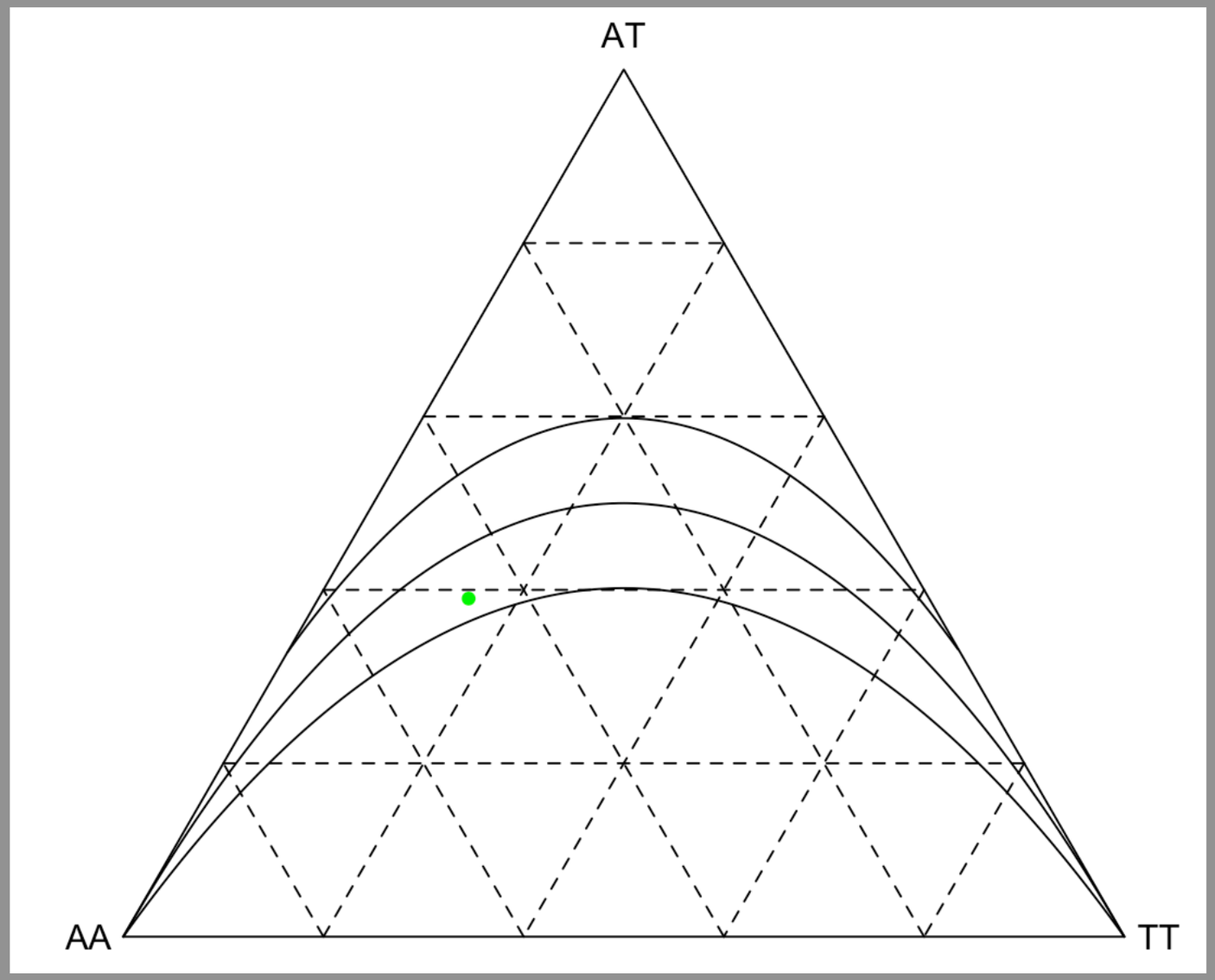
```
> results <- HWChisq(x,verbose=TRUE)
Chi-square test with continuity correction for Hardy-Weinberg
equilibrium
Chi2 = 1.441744 p-value = 0.2298573
```



Gene of interest

Hardy-Weinberg Equilibrium (HWE)

Pearson's χ^2 test - in a ternary plot



Ternary plot for $f_{AA} = 0.46$, $f_{AT} = 0.39$ and $f_{TT} = 15$

```
> library(HardyWeinberg)
> x <- c(46,39,15)
> HWTernaryPlot(x,region=1)
```

Acceptance region for HWE for a Chi Square test without continuity correction

$$2pq - 2pq\sqrt{\chi_1^2(\alpha)/n} \leq f_{AB} \leq 2pq + 2pq\sqrt{\chi_1^2(\alpha)/n}$$

Hardy-Weinberg Equilibrium (HWE)

Statistical tests

- Pearson's χ^2 test.
- **Fisher's exact test** (based on $P(N_{AB} \mid N_A)$).
- Permutation test.
- …

One can use these test not only to test for HWE but for any other expected allele proportions

If the frequencies are very small, see Fisher's exact test

Hardy-Weinberg Equilibrium (HWE)

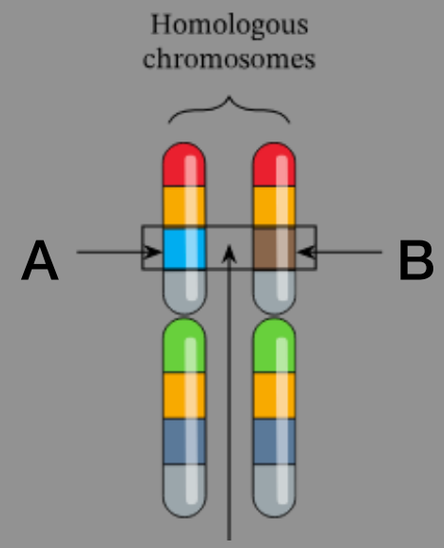
Fisher's exact test

- Fisher's exact test takes into account the number of samples observed n and its counts n_{AA}, n_{AB} and n_{AB} and the p-value can be calculated exactly…rather than relying on an underlying distribution of the parameters.
- Exact non-parametric statistical significant test useful for categorical data when sample sizes are very small.
- Probability of observing a set of categorical values is given by:

	Men	Women	*Row Total*
Studying	***a***	***b***	*a + b*
Non-studying	***c***	***d***	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d (=n)*

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!}$$

- In order to calculate the significance of the observed data, note that the p-value corresponds to sum all probabilities of samples as extreme or more extreme as the one you observed.
- It takes much more CPU than a χ^2 test (use recursion).
- It is conservative.



Gene of interest

Levene, H., 1949. On a matching problem arising in genetics. The annals of mathematical statistics, pp.91-94.

Hardy-Weinberg Equilibrium (HWE)

Fisher's exact test

- Fisher's exact test takes into account the number of samples observed n and its counts n_{AA}, n_{AB} and n_{AB} and the p-value can be calculated exactly…rather than relying on an underlying distribution of the parameters.
- Exact non-parametric statistical significant test useful for categorical data when sample sizes are very small.
- Stevens, Levene, Haldane (Levene 1949) adapted for HWE:

$$P(N_{AA} = n_{AA}, N_{AB} = n_{AB}, N_{BB} = n_{BB}) = \frac{n!}{n_{AA}!n_{AB}!n_{BB}!} \left(p_A^2\right)^{n_{AA}} (2p_A p_B)^{n_{AB}} \left(p_B^2\right)^{n_{BB}}$$

Under the assumption of HWP, the genotypes counts will follow the multinomial distribution with probability vector (p2, 2pq, q2)

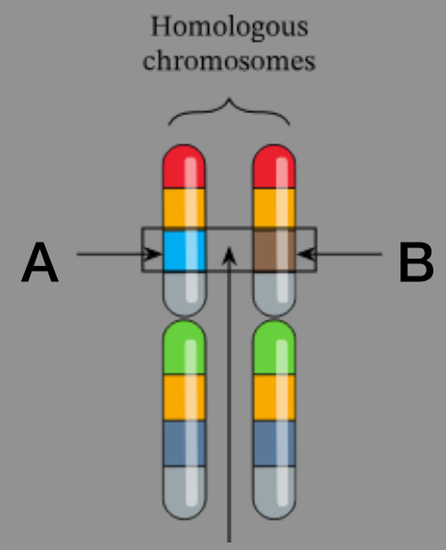
$$P(N_A = n_A) = \frac{2n!}{n_A!n_B!} (p_A)^{n_A} (p_B)^{n_{BB}}$$

Under HWP, all alleles are independent, and the distribution of NA is given by the binomial distribution

$$P(N_{AA}, N_{AB}, N_{BB}|n_A, n_B) = \frac{n_A!n_B!n!2^{n_{AB}}}{\frac{1}{2}(n_A - n_{AB})!n_{AB}!\frac{1}{2}(n_B - n_{AB})!(2n)!}$$

The Stevens-Levene-Haldane distribution is then obtained

- In this case the p-value corresponds to sum all probabilities of samples as extreme or more extreme as the one you observed (there are alternatives).
- It takes much more CPU than a χ^2 test (use recursion).
- It is conservative.



Gene of interest

Levene, H., 1949. On a matching problem arising in genetics. The annals of mathematical statistics, pp.91-94.

Hardy-Weinberg Equilibrium (HWE)

Fisher's exact test - examples in R

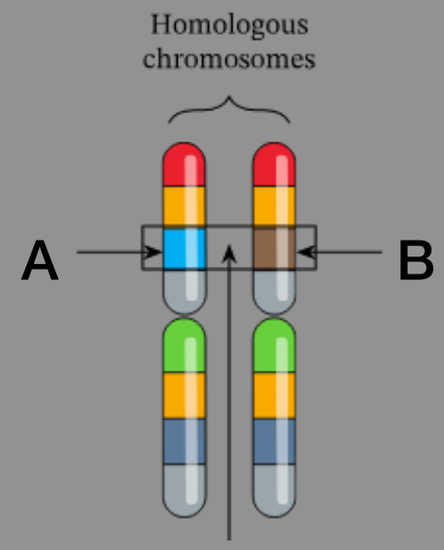
- For an A/T polymorphism with $n_{AA} = 46$, $n_{AT} = 39$ and $n_{TT} = 15$ counts
- Example of the exact test in R

```
> library(HardyWeinberg)
> x <- c(46,39,15)
> HWExact(x,pvaluetype="selome",verbose=TRUE)

Haldane's Exact test for Hardy-Weinberg equilibrium

sample counts: nAA =  46 nAB =  39 nBB =  15

H0: HWE (D==0), H1: D <> 0
D =  -3.0975 p =  0.1852682
```



Gene of interest

Hardy-Weinberg Equilibrium (HWE)

Fisher's exact test - numerical example

- Let A be a rare, disease-predisposing allele with $p_A = 0.025$ (at birth, say).

	f_{AA}	f_{AB}	f_{BB}	p_A
Initial	p^2	$2pq$	q^2	
Population	0.0006	0.0488	0.9506	0.0250

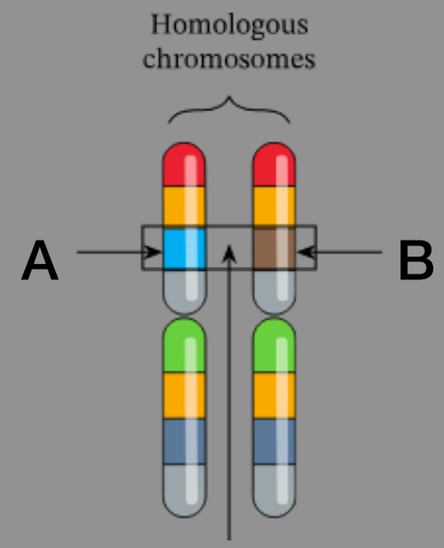
- Then, potentially after many years, genetic variation within each population differs and we obtain:

	f_{AA}	f_{AB}	f_{BB}	p_A
Diseased	0.0128	0.4998	0.4873	0.2627
Non-diseased	0.0001	0.0304	0.9694	0.0153

- Sampling from these distributions (n = 1000), and testing each population for HWP with an exact test:

	AA	AB	BB	Exact p-value
Diseased	11	510	479	≈ 0
Non-diseased	0	19	981	≈ 1

- Disequilibrium observed in cases, but not detected in controls.



Gene of interest

Hardy-Weinberg Equilibrium (HWE)

Statistical tests

- Pearson's χ^2 test.
- Fisher's exact test (based on $P(N_{AB} \mid N_A)$).
- **Permutation test**.
- …

Hardy-Weinberg Equilibrium (HWE)

Permutation test (Monte Carlo scheme)

- The permutation test will estimate the underlying population distribution from where our observations came from.
    1. Compute a test statistic (e.g. χ2, nAB, ...) for the observed data.
    2. Calculate the pseudo-statistic test distribution
        a. Obtain a list of A and B alleles from the observed genotype data (n_{AA}, n_{AB} and n_{BB}).
        b. Permute the alleles and assemble pairs of alleles into genotypes (and note that genotype AB is equivalent to BA).
        c. Compute the test statistic for the permuted data set (pseudo-statistic)
        d. Repeat this N times.
    3. Calculate p-value
        e. Count the number of times the pseudo-statistic is as larger or larger than the value for the observed data (C)
        f. Calculate the p-value as C/N.

- Example of the permutation test in R

SIDE NOTE
The permutation test with alpacas:
https://www.jwilber.me/permutationtest /

```
> x <- c(46,39,15)
> names(x) <- c("AA","AT","TT")
> HWPerm(x)
Permutation test for Hardy-Weinberg equilibrium
Observed statistic: 1.878892   17000 permutations. p-value: 0.1864706
```

Content

Hardy-Weinberg Equilibrium (HWE)

Until here we've responded to: is our genetic marker in HWE?

Now we aim to respond: how far from equilibrium are we?

1. Introduction
2. Graphics & tests
3. Disequilibrium measures
4. Multiple alleles
5. Other cases
    5.1. X chromosome
    5.2. HWE genome-wide
6. Computer exercise

Hardy-Weinberg Equilibrium (HWE)

Measures of (dis)equilibrium

- Several statistics are being used as measures of the degree of disequilibrium:
  - The χ^2 statistic of a Pearson's test for HWE.
  - The p-value of an exact test.
  - The inbreeding coefficient f
  - The theta ratio $\theta = \frac{f_{AB}^2}{f_{AA} f_{BB}}$
  - …

Hardy-Weinberg Equilibrium (HWE)

Measures of (dis)equilibrium - the inbreeding coefficient f

The coefficient of inbreeding of an individual is the probability that two alleles at any locus in an individual are identical by descent from the common ancestor(s) of the two parents.

$$P_{AA} = p_A^2 + p_A p_B f$$
$$P_{AB} = 2 p_A p_B (1 - f)$$
$$P_{BB} = p_B^2 + p_A p_B f$$

It can be shown that:

- $f = 0$: HWE
- $f = 1$: No heterozygotes
- $f < 0$: Heterozygote excess
- $f > 0$: Heterozygote scarcity

$$\frac{-p_m}{1 - p_m} \leq f \leq 1 \text{ with } p_m = min(p_A, p_B)$$

A value lower than 5% is desirable. An inbreeding coefficient of 25% indicates the mating of parents and children. An inbreeding coefficient of 6.25% indicates the mating of first cousins

For a sample data, f can be estimated by ML as $\hat{f} = \frac{4 n_{AA} n_{BB} - n_{AB}^2}{n_A n_B}$

Relates to chi-square statistics $\chi^2 = n\hat{f}^2$

Content

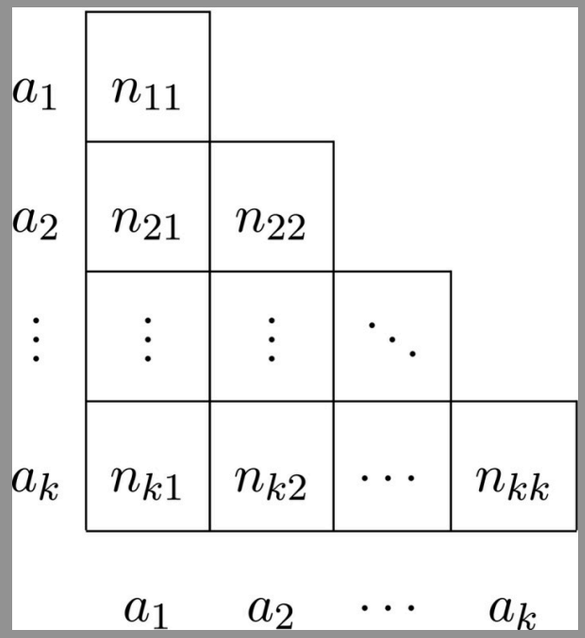
Hardy-Weinberg Equilibrium (HWE)

1. Introduction
2. Graphics & tests
3. Disequilibrium measures
4. Multiple alleles
5. Other cases
    5.1. X chromosome
    5.2. HWE genome-wide
6. Computer exercise

Hardy-Weinberg Equilibrium (HWE)

Multiple alleles

- With many alleles, some are common and many are rare. Consider sample size n where we find a variant with k alleles and n_i represent the sample count of the i^{th} allele (so that $i = 1 \ldots k$).
  - n_{ii} refers to an homozygote $a_i a_i$ and n_{ij} refers to an heterozygote $a_i a_j$.
  - $h = \sum_{i>j} n_{ij}$ is the total heterozygote frequency.



- Asymptotic procedures do not work well with rare alleles (small counts).
- Exact procedures and permutation tests are preferable.
- Computational cost increases
- Exact density for multiple alleles: (Levene 1949)

$$P\ (N_{ij} = n_{ij} | n_1, \ldots, n_k) = \frac{n! 2^h \prod_{i=1}^{k} n_i!}{(2n)! \prod_{i \geq j} n_{ij}!}$$

Hardy-Weinberg Equilibrium (HWE)

Multiple alleles - example with R

Multiple allele testing with exact test and permutation test:

	A	B	C
A	20		
B	31	15	
C	26	12	0

```
> x <- c(AA=20,AB=31,AC=26,BB=15,BC=12,CC=0)
> out <- HWTriExact(x)
Tri-allelic Exact test for HWE (autosomal).
Allele counts: A = 97 B = 73 C = 38
probability of the sample 0.0001122091
p-value =  0.03370688
>
```

```
> x3 <- toTriangular(x)
> out <- HWPerm.mult(x3)
Permutation test for Hardy-Weinberg equilibrium (autosomal).
3 alleles detected.
Observed statistic: 0.0001122091
17000 permutations.
p-value: 0.03405882
```

Content

Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg Equilibrium (HWE)

Other cases

How to test for equilibrium if…

- The variant is X-chromosomal
- The organism studied is tetraploid
- The variant studied has multiple copies
- …

SOURCE: wiki on Gene duplication

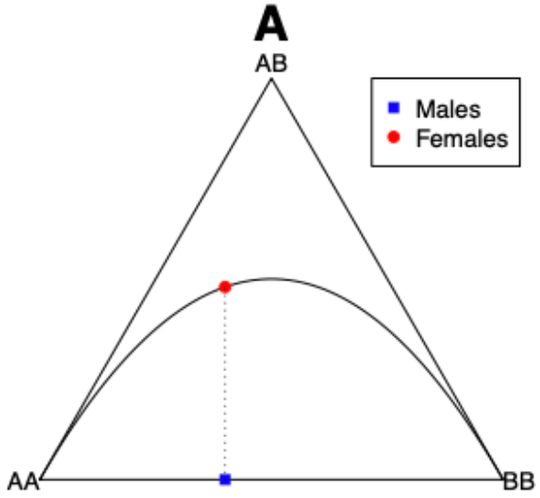
Common oncogene amplifications in human cancers

Cancer type	Associated gene amplifications	Prevalence of amplification in cancer type (percent)
Breast cancer	MYC	20%[39]
	ERBB2 (HER2)	20%[39]
	CCND1 (Cyclin D1)	15–20%[39]
	FGFR1	12%[39]
	FGFR2	12%[39]
Cervical cancer	MYC	25–50%[39]

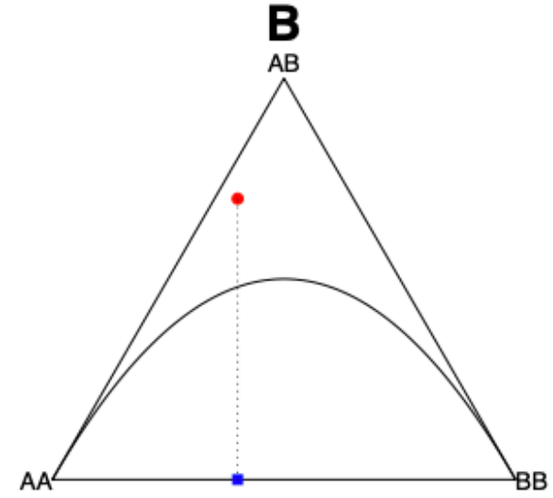
Hardy-Weinberg Equilibrium (HWE)

Other cases - X chromosome

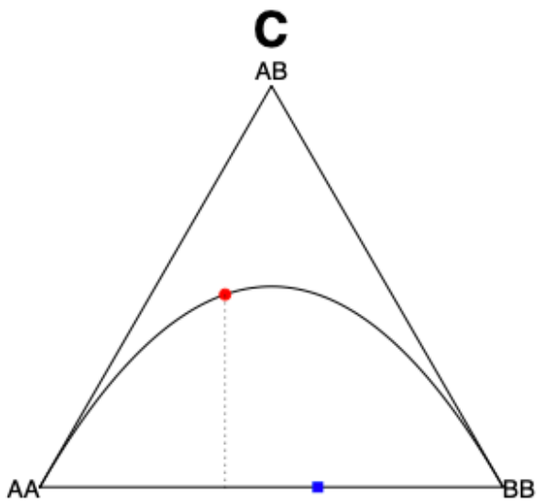
- Genes carried on sex chromosomes, such as the X chromosome of humans, show different inheritance patterns than genes on autosomal (non-sex) chromosomes (sex linkage).
- For a marker on the X chromosome, it can take several generations before HWE is reached.
- A marker on the X chromosome is in HWE if and only if
  - Females occur in the HWP proportions (AA: p2, AB: 2pq BB: q2).
  - Male and female allele frequencies are equal.
- Four scenarios are possible.



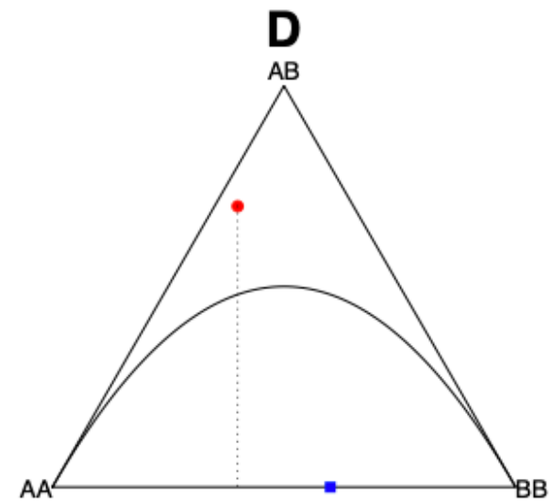
Equilibrium



Disequilibrium due to deviating female genotype frequencies.



Disequilibrium due to nonhomogeneous allele frequencies

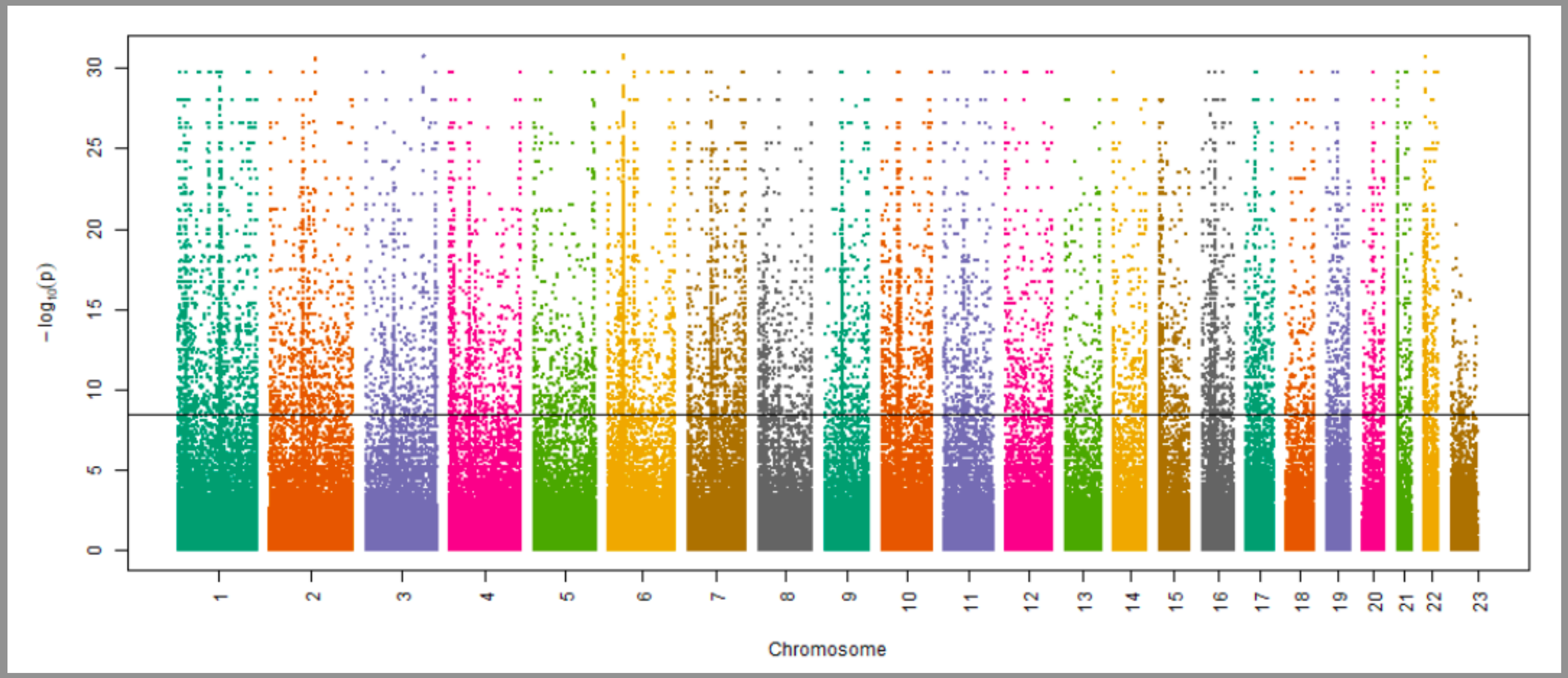


Disequilibrium due to deviating female genotype frequencies and nonhomogeneous allele frequencies.

Graffelman, J. and Weir, B.S., 2016. Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity, 116(6), pp.558-568.

Hardy-Weinberg Equilibrium (HWE)

Other cases - Genome-wide testing for HWE



Manhattan plot of -log10 transformed exact p-values for Hardy-Weinberg equilibrium of 13.4 million SNPs of the CHB sample. The horizontal line corresponds to the Bonferroni significance threshold

Content

Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg Equilibrium (HWE)

R software for studying HWP

- Plink (Purcell, 2007)
- R-package HWEBayes (Wakefield, 2010)
- R-package HardyWeinberg (Graffelman, 2008)
- R-package HWEintrinsic (Venturini, 2011)
- R-package hwde (Maindonald & Johnson, 2011) ...
- …

Hardy-Weinberg Equilibrium (HWE)

References

- Graffelman, J. (2020) Statistical tests for the Hardy-Weinberg equilibrium. Wiley StatsRef: Statistics Reference Online. doi: 10.1002/9781118445112.stat08274.
- Graffelman, J. (2015) Exploring Diallelic Genetic Markers: The HardyWeinberg Package. The Journal of Statistical Software 64(3): 1–23. http://www.jstatsoft.org/v64/i03/paper, doi: 10.18637/jss.v064.i03
- Graffelman, J. and Weir, B.S. (2016) Testing for Hardy-Weinberg equilibrium at bi-allelic genetic markers on the X chromosome. Heredity 116(6) pp. 558–568. doi: 10.1038/hdy.2016.20.
- Hartl, D. L. (1980) Principles of population genetics, Sinauer Associates, Massachusetts,
- Weir, B.S. (1996) Genetic Data Analysis II, Chapter 3, Sinauer Associates, Massachusetts.