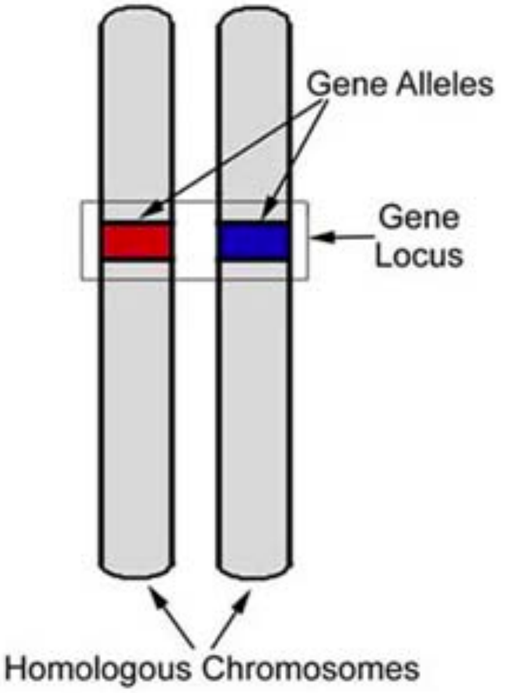
Lecture 3: Genotype and allele frequencies

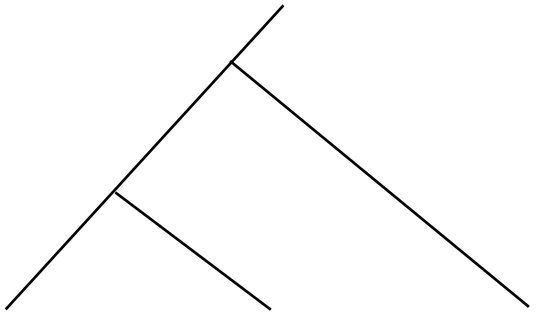
Population genetic PCB4553/6685

Loci and alleles

- **Locus**: specific, fixed position on a chromosome
- A locus may be an entire gene, or a single nucleotide base pair such as A-T.
- **Alleles**: genetic variants segregating in the population at a given locus
- If there are multiple alleles in the population at a locus, we say that this locus is **polymorphic** (this is sometimes referred to as a segregating site)



Genetic variation



Drosophila melanogaster *D. simulans* *D. yakuba*

pos.	con.	a	b	c	d	e	f	g	h	i	j	k	l	a	b	c	d	e	f	a	b	c	d	e	f	g	h	i	j	k	l	NS/S
781	G	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NS
789	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
808	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
816	G	T	T	T	T	-	-	-	-	-	-	-	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
834	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	S
859	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
867	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	A	G	G	G	G	G	G	S
870	C	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
950	G	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
974	G	-	-	-	-	-	-	-	-	-	-	-	-	T	-	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
983	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
1019	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S
1031	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	S
1034	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		C	C	C	C	C	-	-	C	-	C	C	S
1043	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S
1068	C	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
1089	C	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	NS
1101	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	NS
1127	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
1131	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	S
1160	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S

*Columns corresponding to different haplotypes

Measuring genetic variability

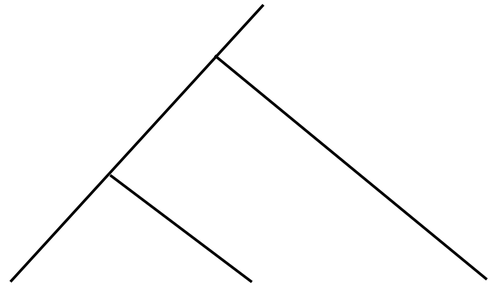
- **Nucleotide diversity** (π): average number of single nucleotide differences between haplotypes chosen at random from a sample

- π for *D. simulans*

$$\pi = \frac{1}{15}\big((2+1+1+1+0)+(3+3+3+2)+(0+0+1)+(0+1)+(1)\big) = 1.2\overline{6}$$

- Normalize by sequence length

$$\pi = 1.2\overline{6}/397 = 0.0032$$



Drosophila melanogaster *D. simulans* *D. yakuba*

pos.	con.	a	b	c	d	e	f	g	h	i	j	k	l	a	b	c	d	e	f	a	b	c	d	e	f	g	h	i	j	k	l	NS/S
781	G	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NS
789	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
808	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
816	G	T	T	T	T	-	-	-	-	-	-	-	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
834	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	S
859	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
867	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	A	G	G	G	G	G	G	S
870	C	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
950	G	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
974	G	-	-	-	-	-	-	-	-	-	-	-	-	T	-	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
983	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
1019	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S
1031	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	S
1034	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		C	C	C	C	C	-	-	C	-	C	C	S
1043	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S
1068	C	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
1089	C	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	NS
1101	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	NS
1127	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
1131	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	S
1160	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S

Nucleotide diversity across species

Measuring genetic variability

- **Allele frequency spectrum**: distribution of the allele frequencies of a given set of loci (often SNPs) in a population or sample
- For every locus, count number of minor allele
- Make histogram with bin_width=1
- Example: allele frequency spectrum is (4, 2, 1, 0, 1)

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8
Sample 1	0	1	0	0	0	0	1	0
Sample 2	1	0	1	0	0	0	1	0
Sample 3	0	1	1	0	0	1	0	0
Sample 4	0	0	0	0	1	0	1	1
Sample 5	0	0	1	0	0	0	1	0
Sample 6	0	0	0	1	0	1	1	0
Total	1	2	3	1	1	2	5	1

Genotype and allele frequency

- We can recover the allele frequencies from genotype frequencies
- Consider a diploid autosomal locus segregating for two alleles (A1 and A2)
- Frequency of A1 in the population:

$$p = \frac{2N_{11} + N_{12}}{2N} = f_{11} + \frac{1}{2} f_{12}.$$

- Can we do the opposite? i.e. calculate genotype frequencies from allele frequencies

Genotype	A1A1	A1A2	A2A2
Number	N11	N12	N22
Frequency	f11	f12	f22

Hardy–Weinberg proportions

- **Relating allele frequencies to genotype frequency**
- By assuming random mating
- Probability of sampling the two alleles are independent with each other
- $P(A_{\text{mom}} = A_2 \,|\, A_{\text{dad}} = A_1) = P(A_{\text{mom}} = A_2)$
- $P(A_1A_2) = P(A_1)P(A_2)$
- HWE is achieved if mating is random with respect to our focal allele
- Multi-allelic loci work in the same manner

			Maternal gamete	
			A1	A2
			p	q
Paternal gamete	A1	p	p^2	pq
	A2	q	qp	q^2

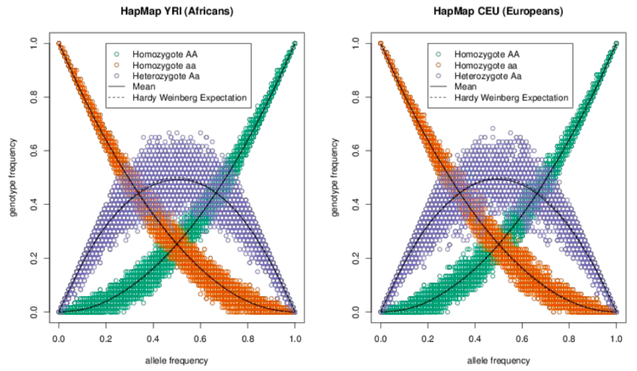


Figure 2.6: Demonstrating Hardy–Weinberg proportions using 10,000 SNPs from the HapMap European (CEU) and African (YRI) populations. Within each of these populations the allele frequency against the frequency of the 3 genotypes; each SNP is represented by 3 different coloured points. The solid lines show the mean genotype frequency. The dashed lines show the predicted genotype frequency from Hardy–Weinberg equilibrium. Code here. Blog post on figure here.

Question 3.

On the coastal islands of British Columbia there is a subspecies of black bear (*Ursus americanus kermodei*, Kermode's bear). Many members of this black bear subspecies are white; they're sometimes called spirit bears. These bears aren't hybrids with polar bears, nor are they albinos. They are homozygotes for a recessive change at the MC1R gene. Individuals who are GG at this SNP are white, while AA and AG individuals are black. Below are the genotype counts for the MC1R polymorphism in a sample of bears from British Columbia's island populations from Ritland *et al.* (2001).

AA	AG	GG
42	24	21

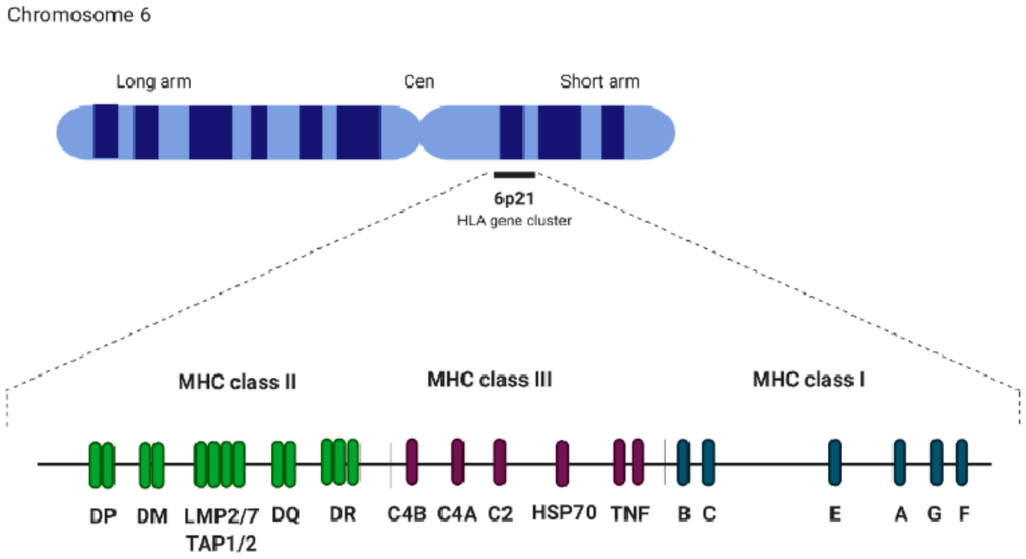
What are the expected frequencies of the three genotypes under HW?

Question 4.

You are investigating a locus with three alleles, A, B, and C, with allele frequencies p_A, p_B, and p_C. What fraction of the population is expected to be homozygotes under Hardy–Weinberg?

Deviations from HWE

- Assortative mating
  - $P(A_1 \mid A_1) > P(A_1)$
- Disassortative
- $P(A_1 \mid A_1) < P(A_1)$
- Sexual selection has been observed in mice (and possibly humans) choosing to mate with females with different MHCs
- Possibly triggered by olfactory capacity to discriminate MHC-mediated odours



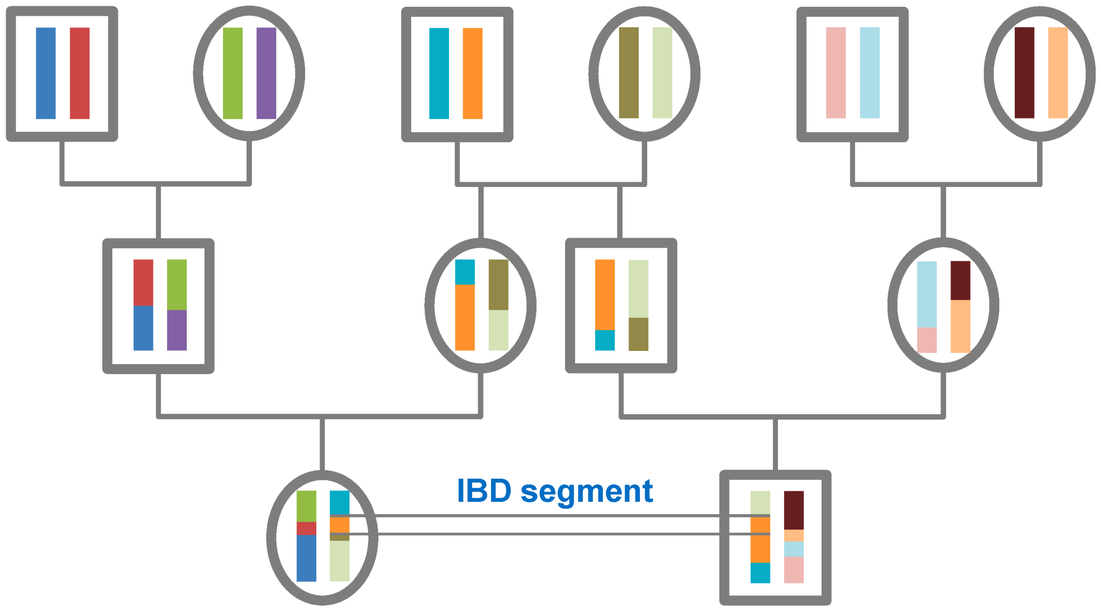
Testing deviations from HWE

- Null hypothesis: probability of observing an allele is independent of the state of the homologous allele
- $$\chi^2 = \sum_k \frac{(\text{Observed frequency}_{ij} - \text{Expected frequency}_{ij})^2}{\text{Expected frequency}_{ij}}$$
- Follows the χ^2-distribution
- With degrees of freedom = $(n_{\text{Rows}} - 1)(n_{\text{Cols}} - 1)$

			Maternal gamete	
			A1	A2
			p	q
Paternal gamete	A1	p	f_{11}	f_{12}
	A2	q	f_{21}	f_{22}

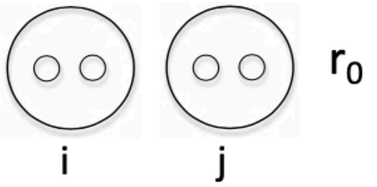
Sharing of alleles among related individuals

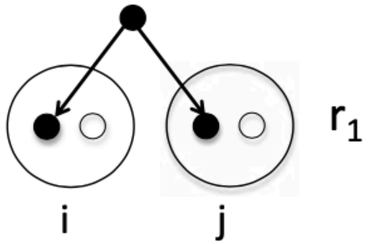
- Population: a group of individuals that are related to each other through common ancestry
- Two alleles are identical by descent (IBD) if they are identical (have the same sequence) due to transmission from a common ancestor in the past
- All individuals in a finite population are related if traced back long enough and will, therefore, share segments of their genomes IBD.
- Different from identical by state (IBS)

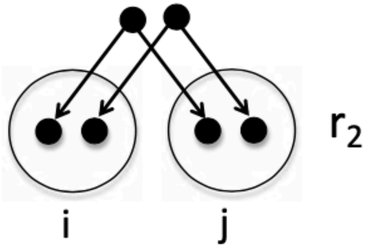


Kinship coefficient

- Kinship coefficient:
  - measure of relatedness
  - gives the probability that two alleles picked at random, one from each of the two different individuals, are identical by descent
- A pair of individuals may share 0, 1, or 2 alleles identical by descent, with probabilities r_0, r_1, r_2
- One way to calculate kinship coefficient is to use these probabilities

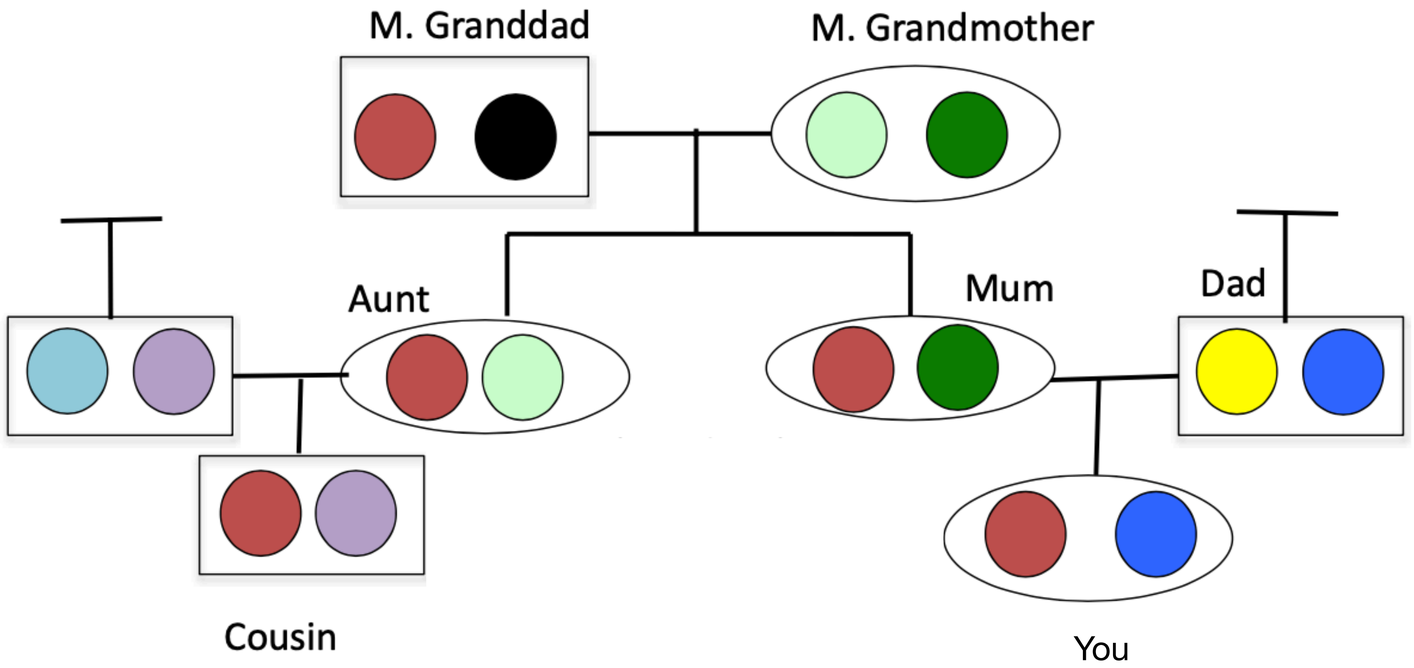






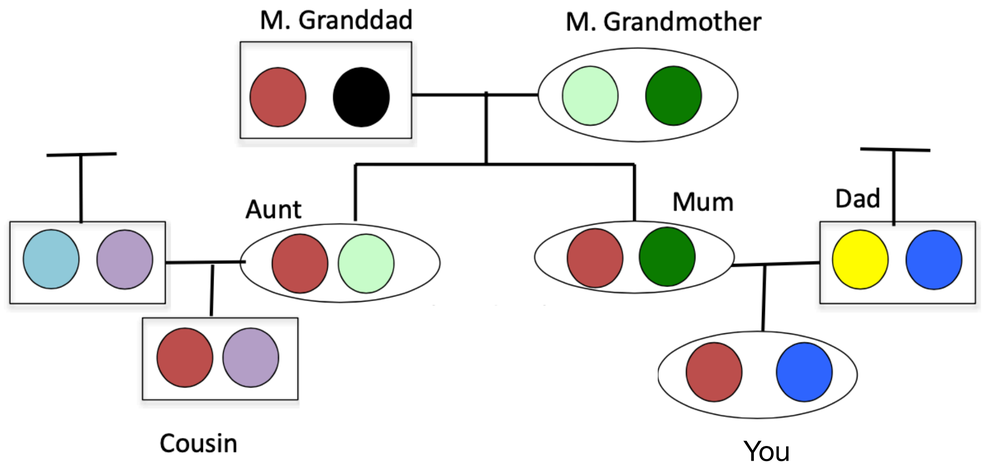
$$
\begin{aligned}
F_{ij} =& \mathbb{P}(\text{I\&J IBD}) \\
=& \mathbb{P}(\text{I\&J IBD}|\ \text{i\&j 0 IBD})\mathbb{P}(\text{i\&j 0 IBD}) \\
&+ \mathbb{P}(\text{I\&J IBD}|\ \text{i\&j 1 IBD})\mathbb{P}(\text{i\&j 1 IBD}) \\
&+ \mathbb{P}(\text{I\&J IBD}|\ \text{i\&j 2 IBD})\mathbb{P}(\text{i\&j 2 IBD}) \\
=& 0 \times r_0 + \frac{1}{4} r_1 + \frac{1}{2} r_2.
\end{aligned}
$$

Genealogical relationships



Allele sharing between relatives

Relationship (i,j)*	$\mathbb{P}$(i&j 0 IBD)	$\mathbb{P}$(i&j 1 IBD)	P(i&j 2 IBD)	$\mathbb{P}$(I&J IBD)
Relationship (i,j)*	r_0	r_1	r_2	F_{ij}
parent–child	0	1	0	$1/4$
full siblings	$1/4$	$1/2$	$1/4$	$1/4$
Monozygotic twins	0	0	1	$1/2$
1^{st} cousins	$3/4$	$1/4$	0	$1/16$



Question 6.

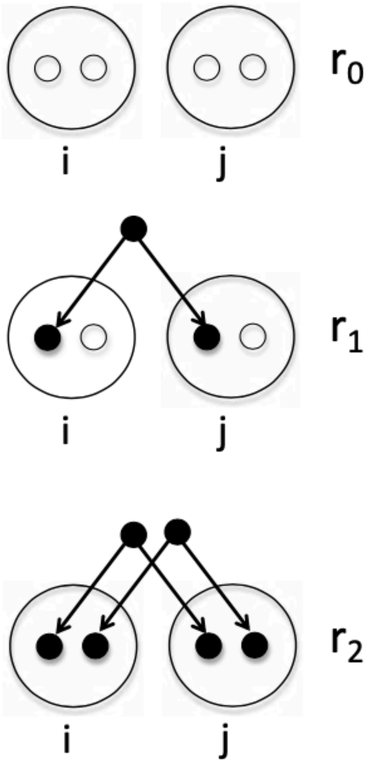
What are r_0, r_1, and r_2 for $^1/_2$ sibs? ($^1/_2$ sibs share one parent but not the other).

Genotype sharing between pairs of individuals

- What is the probability two inds have the identical genotype at a locus (not necessary IBD)
- Apply law of total probability

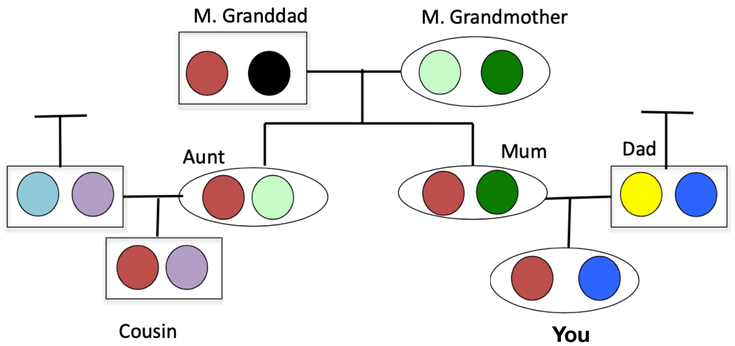
$$\begin{aligned}\mathbb{P}(\text{both } A_1A_1) =& \mathbb{P}(\text{both } A_1A_1|0 \text{ alleles IBD})\mathbb{P}(0 \text{ alleles IBD}) \\ &+ \mathbb{P}(\text{both } A_1A_1|1 \text{ allele IBD})\mathbb{P}(1 \text{ allele IBD}) \\ &+ \mathbb{P}(\text{both } A_1A_1|2 \text{ alleles IBD})\mathbb{P}(2 \text{ alleles IBD})\end{aligned}$$

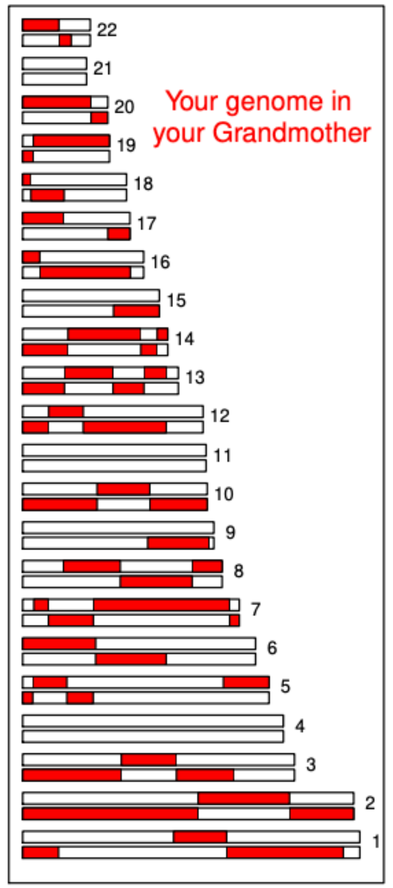
$$\mathbb{P}(\text{both } A_1A_1) = p^4r_0 + p^3r_1 + p^2r_2$$

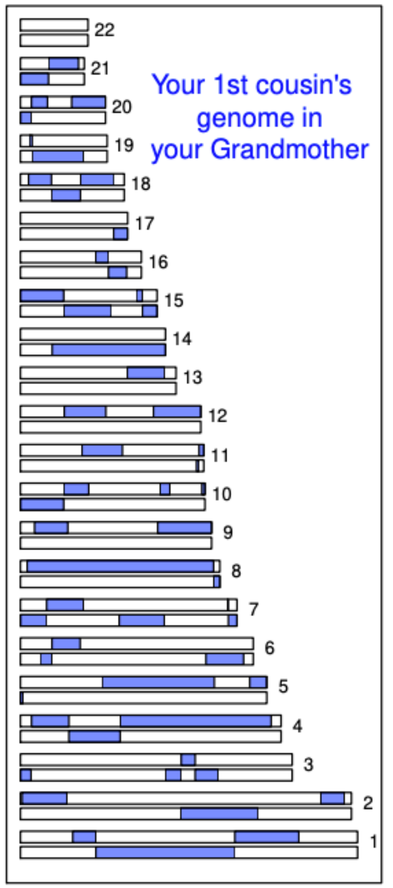


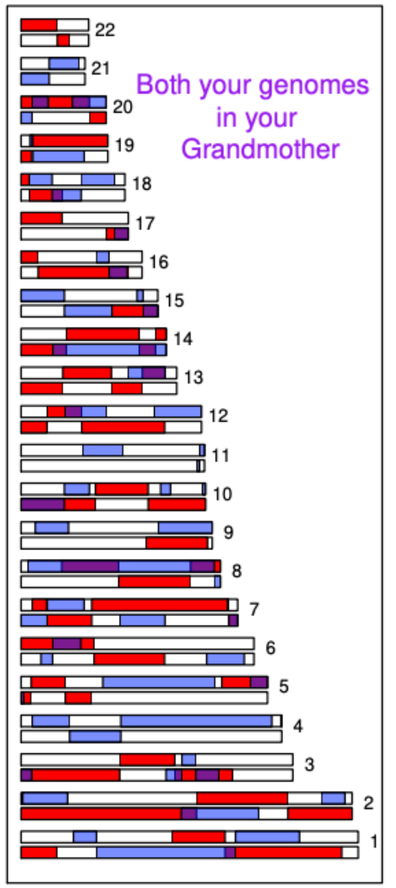
Sharing of genomic blocks among relatives

- Proportion of genomic materials shared among relatives
- Can be derived by looking back in time
- Examples:
  - First cousins share $2 \times (\frac{1}{4})^2 = 1/8$ genomic material
  - Second cousins share $2 \times (\frac{1}{8})^2$ = 1/32
  - Third cousins share $2 \times (\frac{1}{16})^2 = 1/128$





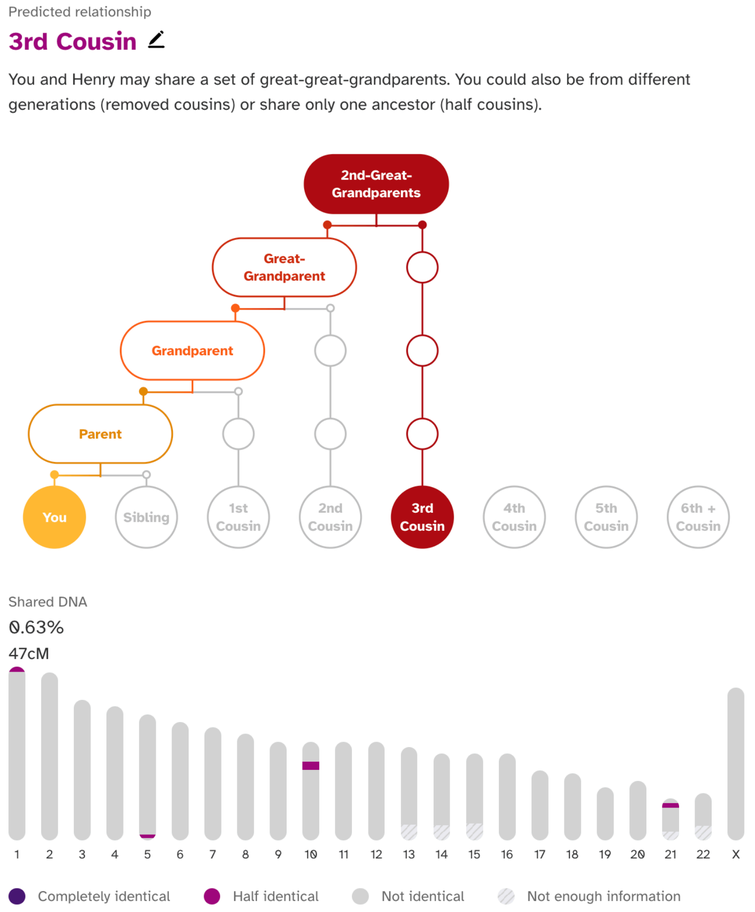




Sharing of genomic blocks among relatives

- High density markers (e.g. SNPs) allows us to identify shared genomic material between relatives in the form of genomic blocks
- We can identify shared blocks by looking for unusually long stretches of the genome where two individuals are never homozygous for different alleles.
- Applications:
  - identify unknown relatives
  - estimate pairwise relationship

“To identify your DNA relatives, we use an algorithm that finds segments of your DNA that are perfect matches to the same DNA segments in other 23andMe customers. When these segments are sufficiently long, we infer that they were likely inherited from a recent shared ancestor… Our algorithm searches for these segments all across your genome, so we can identify DNA relatives on any branch of your family tree.“ - 23andme



Inbreeding

- Inbreeding: mating occurs between individuals that are more closely related to each other than two random individuals drawn from some reference population.

- Genotype frequency under inbreeding

$$\mathbb{P}(A_1A_2) = \mathbb{P}(A_1A_2|\text{I \& J not IBD})\mathbb{P}(\text{I \& J not IBD}) = 2pq(1-F_{ij})$$

$$\begin{aligned} P(A_1A_1) &= \mathbb{P}(A_1A_1|\text{I \& J not IBD})\mathbb{P}(\text{I\&J not IBD}) + \mathbb{P}(A_1A_1|\text{I\&J IBD})\mathbb{P}(\text{I\&J IBD}) \\ &= p^2(1-F_{ij}) + pF_{ij}. \end{aligned} \tag{2.10}$$

- Generalized Hardy-Weinberg

f_{11}	f_{12}	f_{22}
$(1-F)p^2+Fp$	$(1-F)2pq$	$(1-F)q^2+Fq$

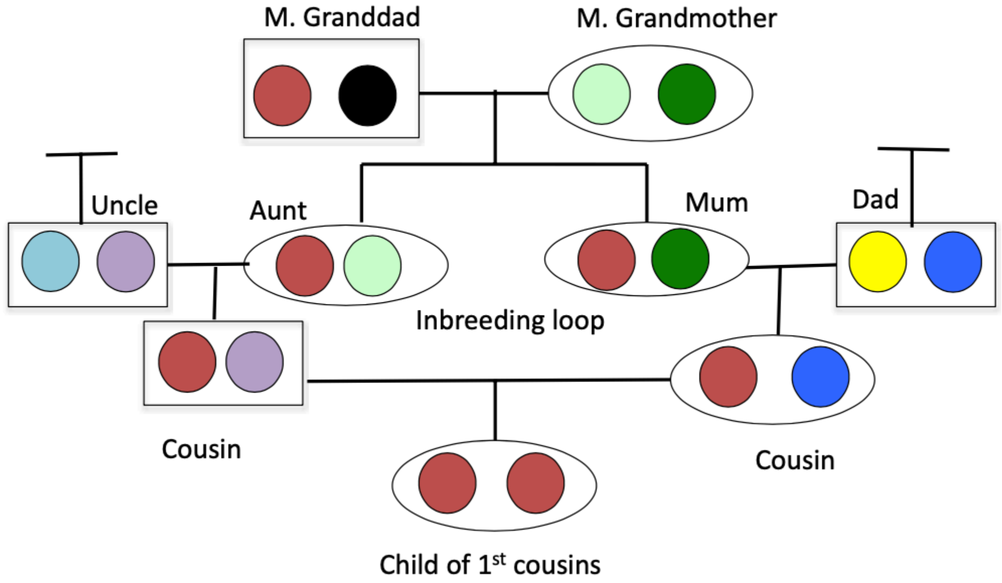
Question 9.

The frequency of the A_1 allele is p at a biallelic locus. Assume that our population is randomly mating and that the genotype frequencies in the population follow from HW. We select two individuals at random to mate from this population. We then mate the children from this cross. What is the probability that the child from this full sib-mating is homozygous?

Calculating inbreeding coefficient from a pedigree

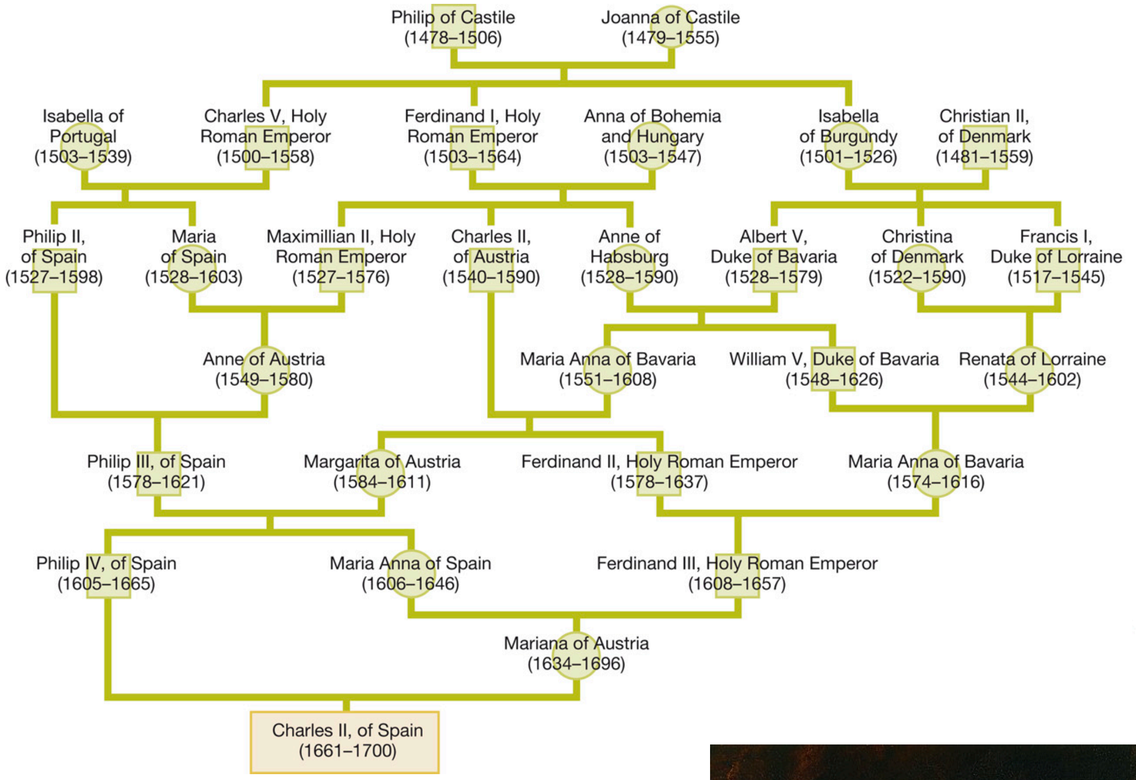
- Inbreeding coefficient of an individual = Kinship coefficient of the parents
- Assume individuals B and C share K common ancestor
- n_i = the total number of individuals in the chain from B to C via ancestor i

$$F = \sum_{i=1}^{K} \frac{1}{2^{n_i}} (1 + f_{A_i})$$



Extreme case of inbreeding

- How many inbreeding loops the lead to Charles II of Spain?
- Inbreeding coefficient =0.254, equivalent to full-sib mating



https://blog.23andme.com/articles/inbreeding-doomed-habsburg

Calculating inbreeding coefficient from genetics data

- Calculate f_{11}, f_{12}, f_{22} using genetic markers
- Solve for F using $f_{12} = (1 - F)2pq$
- $F = 1 - \frac{f_{12}}{2pq}$
- $F = 1 - \frac{H_O}{H_E}$
- Maybe averaged over many markers

f_{11}	f_{12}	f_{22}
$(1-F)p^2 + Fp$	$(1-F)2pq$	$(1-F)q^2 + Fq$

Question 10.

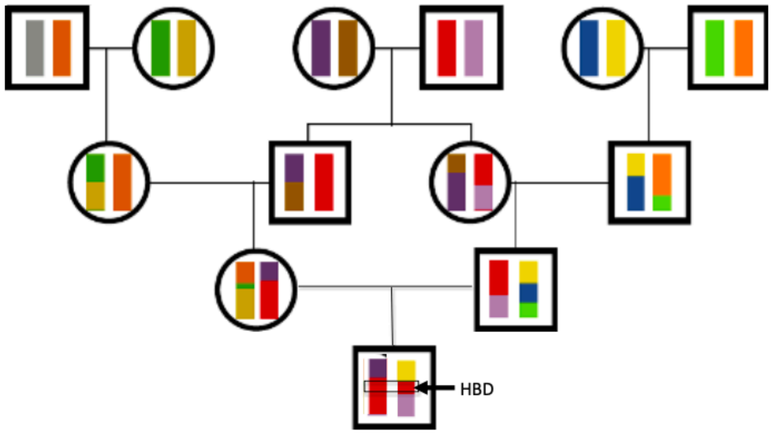
Suppose the following genotype frequencies were observed for an esterase locus in a population of *Drosophila* (A denotes the "fast" allele and B denotes the "slow" allele):

AA	AB	BB
0.6	0.2	0.2

What is the estimate of the inbreeding coefficient at the esterase locus?

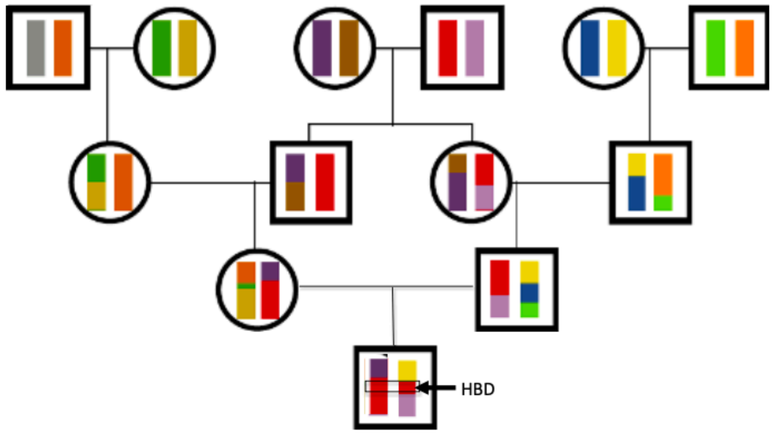
Genomic blocks of homozygosity due to inbreeding

- Close relatives are expected to share alleles IBD in large genomic blocks
- Inbred offspring will have homozygous genomic blocks due to IBD on the two homologous chromosomes (autozygous segment)
- Proportion of autozygous segment = inbreeding coefficient
- Example: child of first-cousin mating have 1/16 of their genome being autozygous



Genomic blocks of homozygosity due to inbreeding

- We can detect these blocks by looking for unusually long genomic runs of homozygosity (ROH) sites in an individual's genome
- Can be used to estimate the inbreeding coefficient of an individual
  - F_{ROH} = Proportion of an individual's genome that falls in such ROH regions



Genomic blocks of homozygosity due to inbreeding

- Individuals with multiple inbreeding loops in their family tree can have a high inbreeding coefficient due to the combined effect of many small blocks of autozygosity
- Distribution of ROH lengths reveals inbreeding history
- Example:
  - *English bulldogs* have had long history of inbreeding as they have many small blocks
  - *Doberman Pinschers* have a lot of recent inbreeding as their autozygosity is contained in long blocks relatively unbroken by recombination.