# POP SUBDIVISION

# DEMOGRAPHY INFERENCE; ILS

# LINKAGE AND LD

# TODAY

- Lecture recap:
  - Wright-Fisher + migration; Wahlund effect; coalescence + migration; divergence
  - genetic linkage; linkage disequilibrium
- Group work: pen-and-paper exercises

# FRIDAY SNEAK PEEK

- We continue the exercises (not enough time for all today) – but in R!

- Discussion – same as today

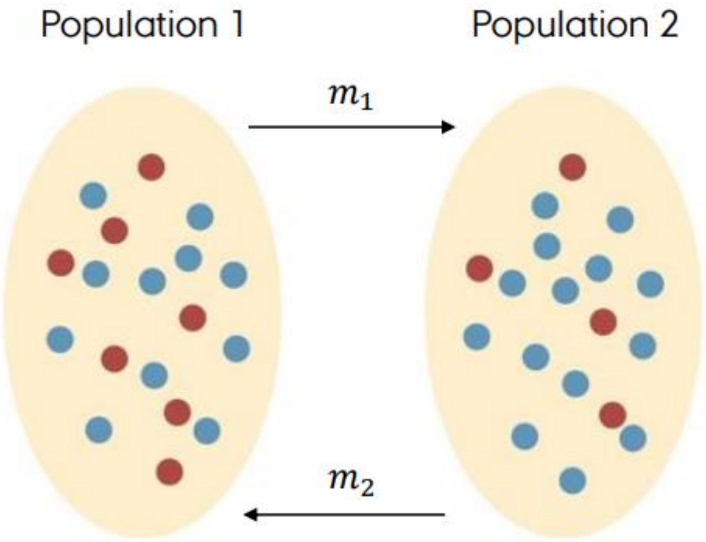
- Menti quiz for the week

# POPULATION SUBDIVISION

- Wright-Fisher model with migration

$$E[f_{A1}(t+1)] = (1-m_1)f_{A1}(t) + m_2 f_{A2}(t)$$

$$E[f_{A2}(t+1)] = (1-m_2)f_{A2}(t) + m_1 f_{A1}(t)$$



- Migration equilibrium

$$E[f_{A1}(t+1)] = f_{A1}(t) \Rightarrow f_{A1}(t) = f_{A2}(t)$$

# POPULATION SUBDIVISION

- Wahlund effect

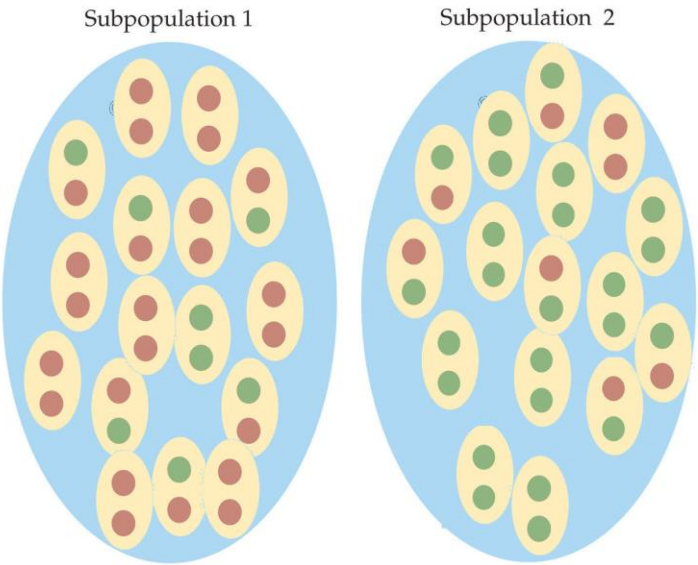
$$f_A = \frac{f_{A1} + f_{A2}}{2} \qquad (4.2)$$

$$H_S = \frac{2f_{A1}(1-f_{A1}) + 2f_{A2}(1-f_{A2})}{2} = f_{A1}(1-f_{A1}) + f_{A2}(1-f_{A2}) \qquad (4.3)$$

$$H_T = 2\frac{f_{A1} + f_{A2}}{2}\left(1 - \frac{f_{A1} + f_{A2}}{2}\right) \qquad (4.4)$$

if $f_{A1} \neq f_{A2}$, then $H_T > H_S$

how about if $f_{A1} = f_{A2}$ ?



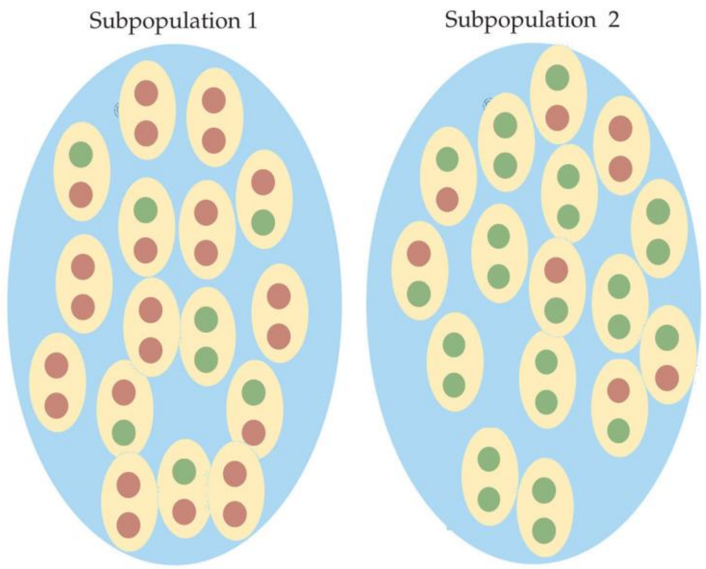
# POPULATION SUBDIVISION

- Wahlund effect

$$f_A = \frac{f_{A1} + f_{A2}}{2} \qquad (4.2)$$

$$H_S = \frac{2f_{A1}(1-f_{A1}) + 2f_{A2}(1-f_{A2})}{2} = f_{A1}(1-f_{A1}) + f_{A2}(1-f_{A2}) \qquad (4.3)$$

$$H_T = 2\frac{f_{A1} + f_{A2}}{2}\left(1 - \frac{f_{A1} + f_{A2}}{2}\right) \qquad (4.4)$$



if $f_{A1} \neq f_{A2}$, then $H_T > H_S$

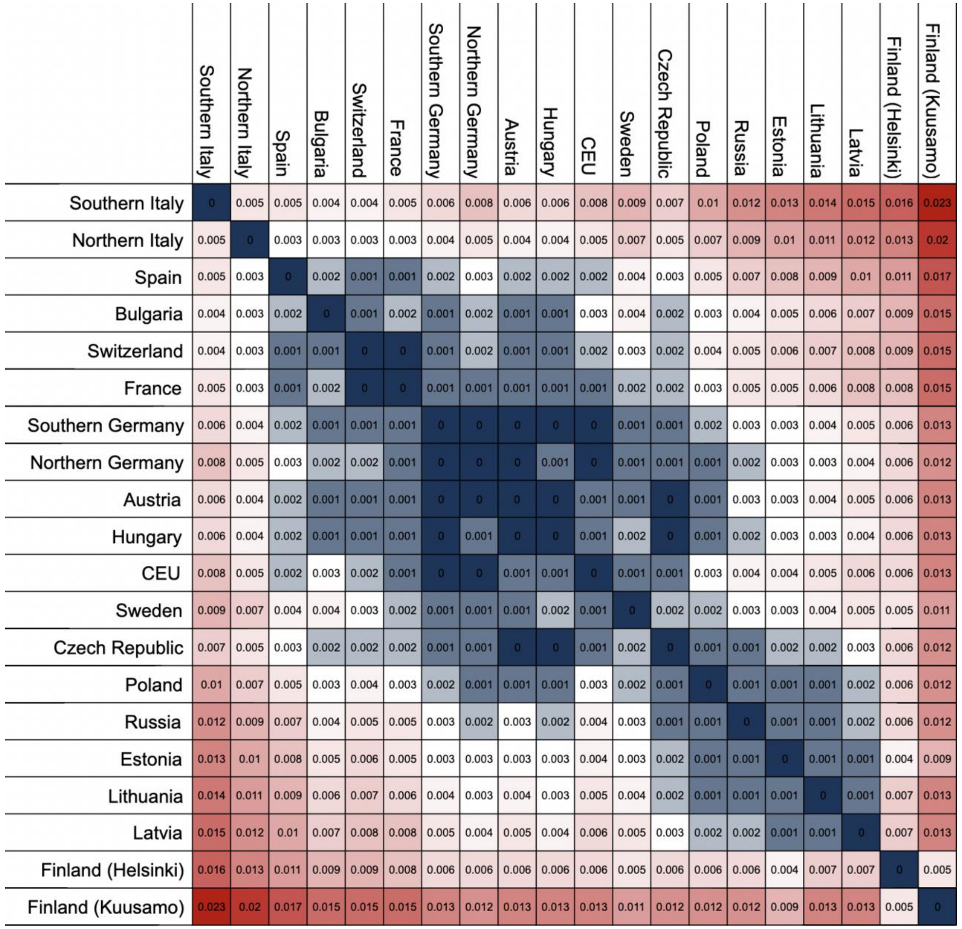
how about if f$_{A1}$ = f$_{A2}$ ?   $H_T = H_S$

# POPULATION SUBDIVISION

- Fixation index

$$F_{ST} = \frac{H_T - H_S}{H_T} \qquad (4.6)$$

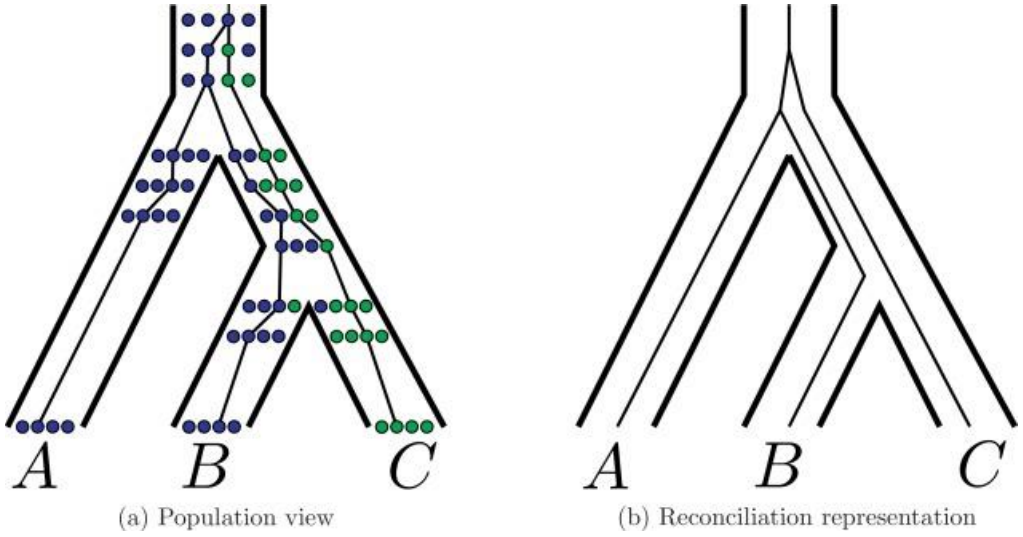
- quantifies population differentiation



| | Southern Italy | Northern Italy | Spain | Bulgaria | Switzerland | France | Southern Germany | Northern Germany | Austria | Hungary | CEU | Sweden | Czech Republic | Poland | Russia | Estonia | Lithuania | Latvia | Finland (Helsinki) | Finland (Kuusamo) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Southern Italy | 0 | 0.005 | 0.005 | 0.004 | 0.004 | 0.005 | 0.006 | 0.008 | 0.006 | 0.006 | 0.008 | 0.009 | 0.007 | 0.01 | 0.012 | 0.013 | 0.014 | 0.015 | 0.016 | 0.023 |
| Northern Italy | 0.005 | 0 | 0.003 | 0.003 | 0.003 | 0.003 | 0.004 | 0.005 | 0.004 | 0.004 | 0.005 | 0.007 | 0.005 | 0.007 | 0.009 | 0.01 | 0.011 | 0.012 | 0.013 | 0.02 |
| Spain | 0.005 | 0.003 | 0 | 0.002 | 0.001 | 0.001 | 0.002 | 0.003 | 0.002 | 0.002 | 0.002 | 0.004 | 0.003 | 0.005 | 0.007 | 0.008 | 0.009 | 0.01 | 0.011 | 0.017 |
| Bulgaria | 0.004 | 0.003 | 0.002 | 0 | 0.001 | 0.002 | 0.001 | 0.002 | 0.001 | 0.001 | 0.003 | 0.004 | 0.002 | 0.003 | 0.004 | 0.005 | 0.006 | 0.007 | 0.009 | 0.015 |
| Switzerland | 0.004 | 0.003 | 0.001 | 0.001 | 0 | 0 | 0.001 | 0.002 | 0.001 | 0.001 | 0.002 | 0.003 | 0.002 | 0.004 | 0.005 | 0.006 | 0.007 | 0.008 | 0.009 | 0.015 |
| France | 0.005 | 0.003 | 0.001 | 0.002 | 0 | 0 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 | 0.003 | 0.005 | 0.005 | 0.006 | 0.008 | 0.008 | 0.015 |
| Southern Germany | 0.006 | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0 | 0 | 0 | 0 | 0 | 0.001 | 0.001 | 0.002 | 0.003 | 0.003 | 0.004 | 0.005 | 0.006 | 0.013 |
| Northern Germany | 0.008 | 0.005 | 0.003 | 0.002 | 0.002 | 0.001 | 0 | 0 | 0 | 0.001 | 0 | 0.001 | 0.001 | 0.001 | 0.002 | 0.003 | 0.003 | 0.004 | 0.006 | 0.012 |
| Austria | 0.006 | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0 | 0 | 0 | 0 | 0.001 | 0.001 | 0 | 0.001 | 0.003 | 0.003 | 0.004 | 0.005 | 0.006 | 0.013 |
| Hungary | 0.006 | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0 | 0.001 | 0 | 0 | 0.001 | 0.002 | 0 | 0.001 | 0.002 | 0.003 | 0.003 | 0.004 | 0.006 | 0.013 |
| CEU | 0.008 | 0.005 | 0.002 | 0.003 | 0.002 | 0.001 | 0 | 0 | 0.001 | 0.001 | 0 | 0.001 | 0.001 | 0.003 | 0.004 | 0.004 | 0.005 | 0.006 | 0.006 | 0.013 |
| Sweden | 0.009 | 0.007 | 0.004 | 0.004 | 0.003 | 0.002 | 0.001 | 0.001 | 0.001 | 0.002 | 0.001 | 0 | 0.002 | 0.002 | 0.003 | 0.003 | 0.004 | 0.005 | 0.005 | 0.011 |
| Czech Republic | 0.007 | 0.005 | 0.003 | 0.002 | 0.002 | 0.002 | 0.001 | 0.001 | 0 | 0 | 0.001 | 0.002 | 0 | 0.001 | 0.001 | 0.002 | 0.002 | 0.003 | 0.006 | 0.012 |
| Poland | 0.01 | 0.007 | 0.005 | 0.003 | 0.004 | 0.003 | 0.002 | 0.001 | 0.001 | 0.001 | 0.003 | 0.002 | 0.001 | 0 | 0.001 | 0.001 | 0.001 | 0.002 | 0.006 | 0.012 |
| Russia | 0.012 | 0.009 | 0.007 | 0.004 | 0.005 | 0.005 | 0.003 | 0.002 | 0.003 | 0.002 | 0.004 | 0.003 | 0.001 | 0.001 | 0 | 0.001 | 0.001 | 0.002 | 0.006 | 0.012 |
| Estonia | 0.013 | 0.01 | 0.008 | 0.005 | 0.006 | 0.005 | 0.003 | 0.003 | 0.003 | 0.003 | 0.004 | 0.003 | 0.002 | 0.001 | 0.001 | 0 | 0.001 | 0.001 | 0.004 | 0.009 |
| Lithuania | 0.014 | 0.011 | 0.009 | 0.006 | 0.007 | 0.006 | 0.004 | 0.003 | 0.004 | 0.003 | 0.005 | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0 | 0.001 | 0.007 | 0.013 |
| Latvia | 0.015 | 0.012 | 0.01 | 0.007 | 0.008 | 0.008 | 0.005 | 0.004 | 0.005 | 0.004 | 0.006 | 0.005 | 0.003 | 0.002 | 0.002 | 0.001 | 0.001 | 0 | 0.007 | 0.013 |
| Finland (Helsinki) | 0.016 | 0.013 | 0.011 | 0.009 | 0.009 | 0.008 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.006 | 0.006 | 0.006 | 0.004 | 0.007 | 0.007 | 0 | 0.005 |
| Finland (Kuusamo) | 0.023 | 0.02 | 0.017 | 0.015 | 0.015 | 0.015 | 0.013 | 0.012 | 0.013 | 0.013 | 0.013 | 0.011 | 0.012 | 0.012 | 0.012 | 0.009 | 0.013 | 0.013 | 0.005 | 0 |

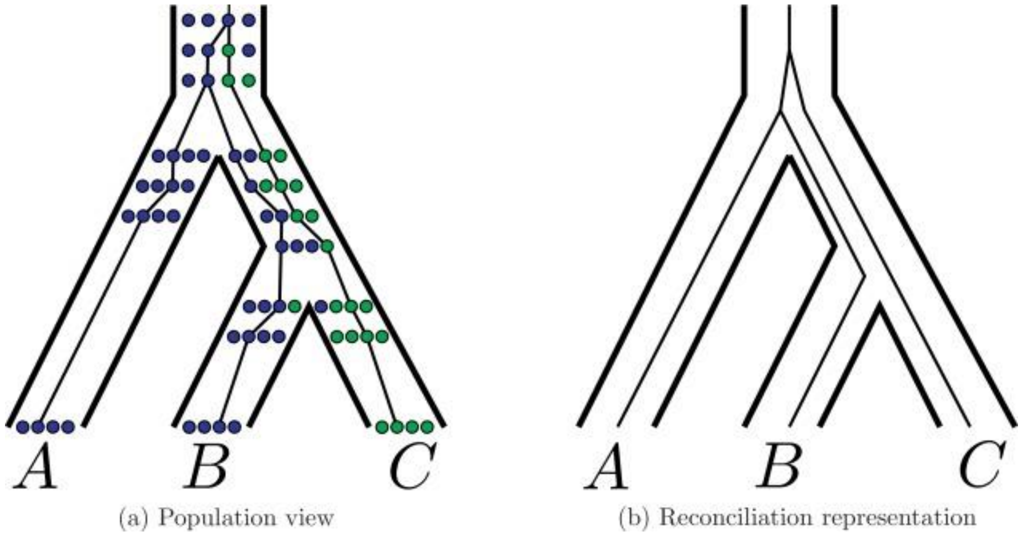
# DEMOGRAPHY INFERENCE

- ILS
    - interspecific coalescent for a given gene is deeper than speciation event
    - Why?



# DEMOGRAPHY INFERENCE

- ILS
  - interspecific coalescent for a given gene is deeper than speciation event
  - Why?
    - Drift
    - Selection
    - Interspecific gene flow
      - (hybridisation)



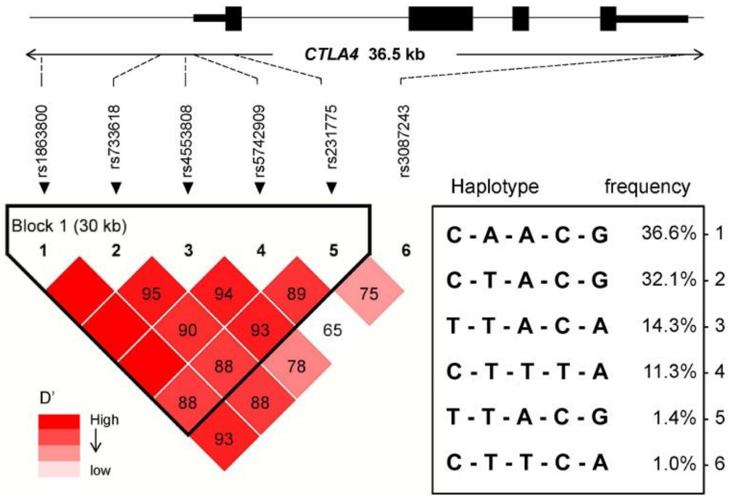
(a) Population view

(b) Reconciliation representation

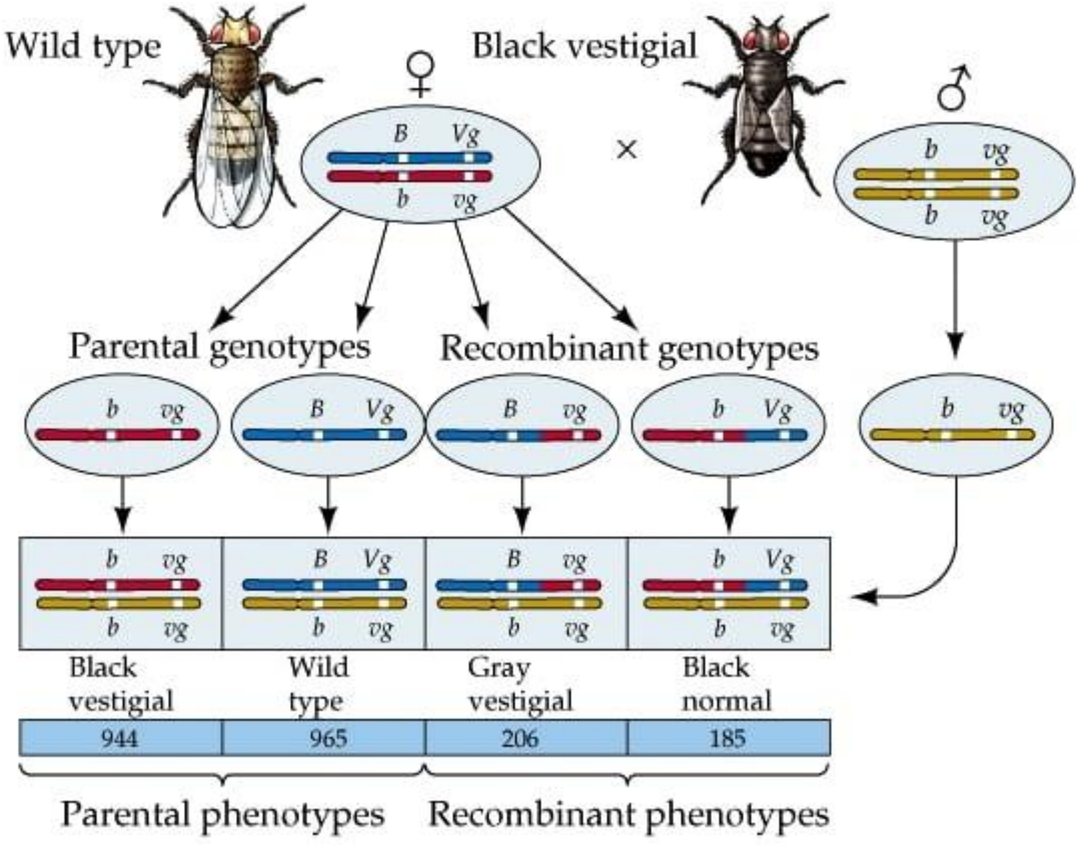
# LINKAGE

- Genetic linkage -> haplotypes

  - Physical association of variants on chromosomes translates into haplotype formation



| Haplotype | frequency | |
|---|---|---|
| C - A - A - C - G | 36.6% | 1 |
| C - T - A - C - G | 32.1% | 2 |
| T - T - A - C - A | 14.3% | 3 |
| C - T - T - T - A | 11.3% | 4 |
| T - T - A - C - G | 1.4% | 5 |
| C - T - T - C - A | 1.0% | 6 |



© 2001 Sinauer Associates, Inc.

# LINKAGE

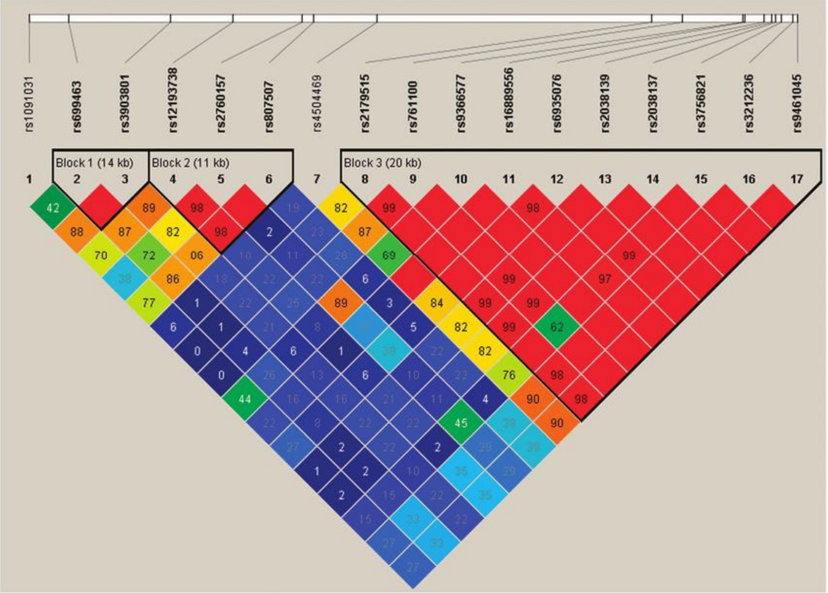
- LD

$$D_{AB} = f_{AB} - f_A f_B \qquad (6.1)$$

$$D_{Ab} = f_{Ab} - f_A f_b$$

$$D_{aB} = f_{aB} - f_a f_B \qquad (6.2)$$

$$D_{ab} = f_{ab} - f_a f_b$$

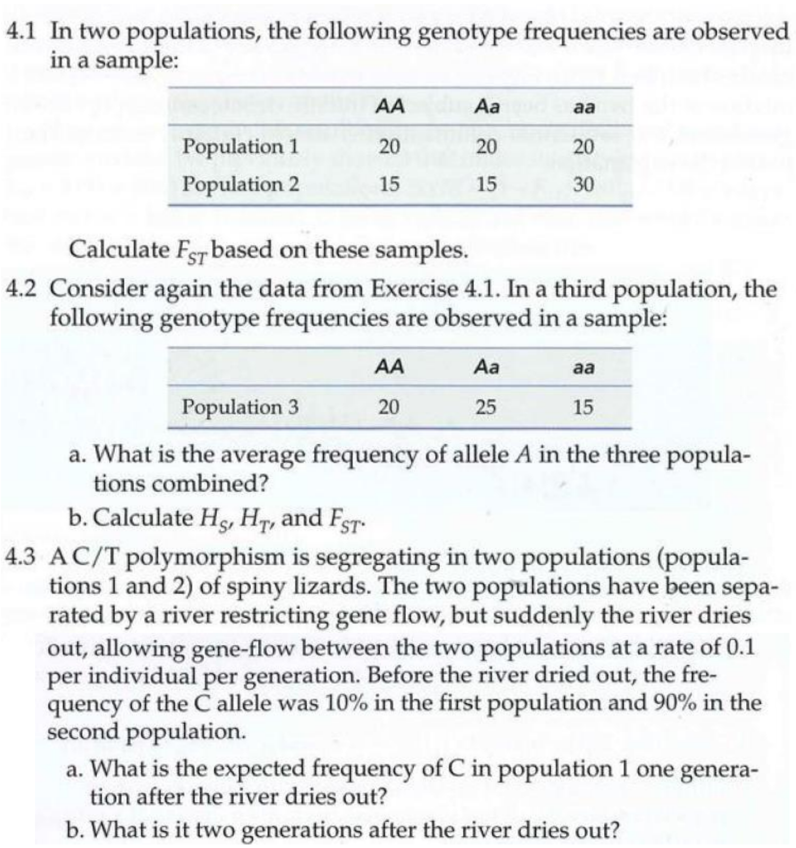


$$\boxed{\begin{array}{l}\text{If } D > 0\text{, then } D \le \min(f_A, f_b, f_a, f_B) \\ \text{If } D < 0\text{, then } -D \le \min(f_A, f_b, f_a, f_B)\end{array}} \qquad \begin{array}{l}(6.4) \\ (6.5)\end{array}$$

$$D(t+1) = (1-r)D(t) \qquad (6.8)$$

$$D(t) = (1-r)^t D(0) \qquad (6.9)$$

# EXERCISE DISCUSSION – CH4



4.1 In two populations, the following genotype frequencies are observed in a sample:

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| Population 1 | 20 | 20 | 20 |
| Population 2 | 15 | 15 | 30 |

Calculate $F_{ST}$ based on these samples.

4.2 Consider again the data from Exercise 4.1. In a third population, the following genotype frequencies are observed in a sample:

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| Population 3 | 20 | 25 | 15 |

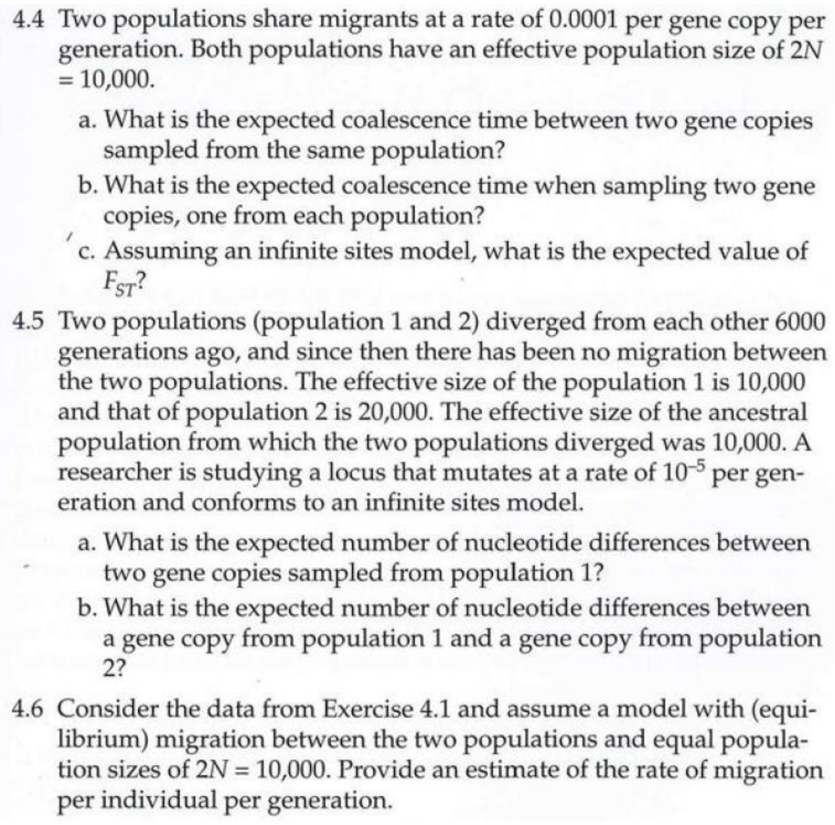
a. What is the average frequency of allele *A* in the three populations combined?

b. Calculate $H_S$, $H_T$, and $F_{ST}$.

4.3 A C/T polymorphism is segregating in two populations (populations 1 and 2) of spiny lizards. The two populations have been separated by a river restricting gene flow, but suddenly the river dries out, allowing gene-flow between the two populations at a rate of 0.1 per individual per generation. Before the river dried out, the frequency of the C allele was 10% in the first population and 90% in the second population.

a. What is the expected frequency of C in population 1 one generation after the river dries out?

b. What is it two generations after the river dries out?



4.4 Two populations share migrants at a rate of 0.0001 per gene copy per generation. Both populations have an effective population size of $2N = 10{,}000$.

a. What is the expected coalescence time between two gene copies sampled from the same population?

b. What is the expected coalescence time when sampling two gene copies, one from each population?

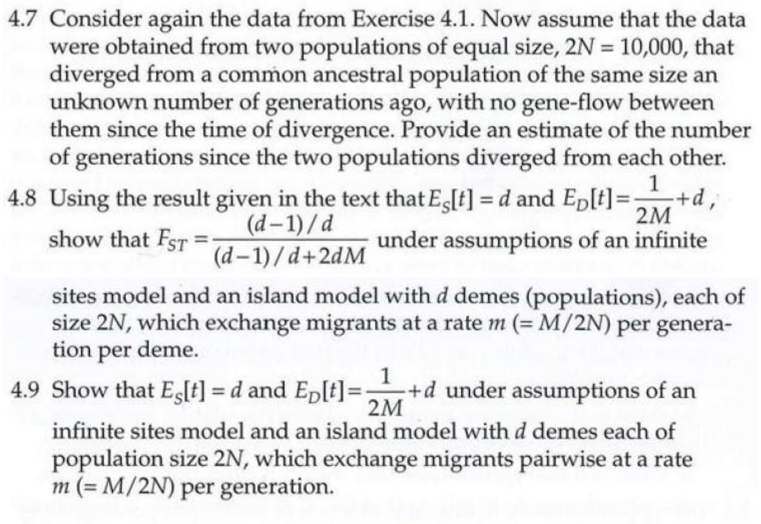
c. Assuming an infinite sites model, what is the expected value of $F_{ST}$?

4.5 Two populations (population 1 and 2) diverged from each other 6000 generations ago, and since then there has been no migration between the two populations. The effective size of the population 1 is 10,000 and that of population 2 is 20,000. The effective size of the ancestral population from which the two populations diverged was 10,000. A researcher is studying a locus that mutates at a rate of $10^{-5}$ per generation and conforms to an infinite sites model.

a. What is the expected number of nucleotide differences between two gene copies sampled from population 1?

b. What is the expected number of nucleotide differences between a gene copy from population 1 and a gene copy from population 2?

4.6 Consider the data from Exercise 4.1 and assume a model with (equilibrium) migration between the two populations and equal population sizes of $2N = 10{,}000$. Provide an estimate of the rate of migration per individual per generation.

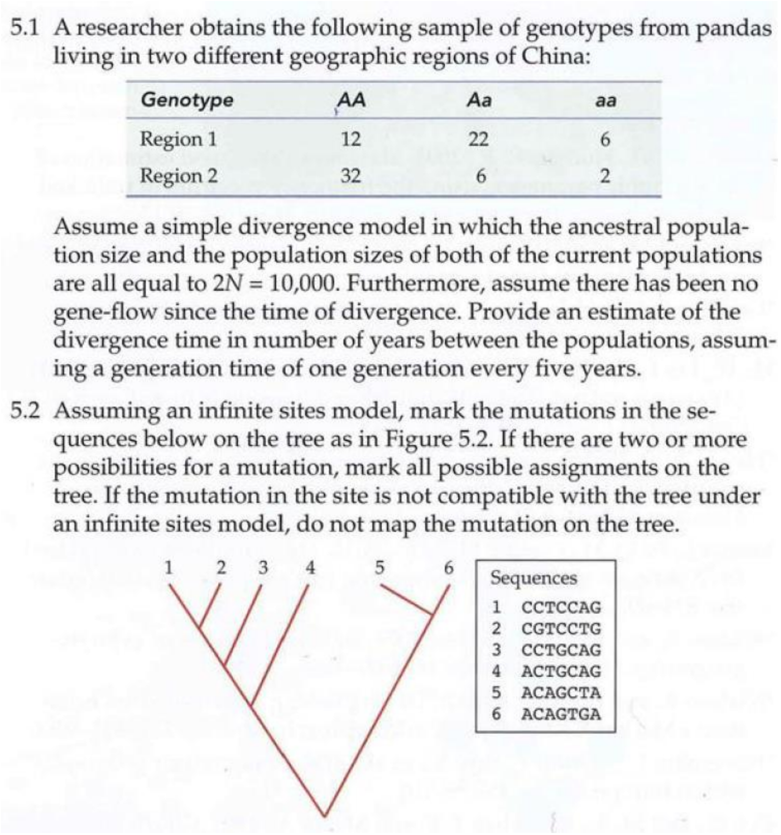


4.7 Consider again the data from Exercise 4.1. Now assume that the data were obtained from two populations of equal size, $2N = 10{,}000$, that diverged from a common ancestral population of the same size an unknown number of generations ago, with no gene-flow between them since the time of divergence. Provide an estimate of the number of generations since the two populations diverged from each other.

4.8 Using the result given in the text that $E_S[t] = d$ and $E_D[t] = \frac{1}{2M} + d$, show that $F_{ST} = \frac{(d-1)/d}{(d-1)/d + 2dM}$ under assumptions of an infinite sites model and an island model with *d* demes (populations), each of size 2*N*, which exchange migrants at a rate $m\ (= M/2N)$ per generation per deme.

4.9 Show that $E_S[t] = d$ and $E_D[t] = \frac{1}{2M} + d$ under assumptions of an infinite sites model and an island model with *d* demes each of population size 2*N*, which exchange migrants pairwise at a rate $m\ (= M/2N)$ per generation.

# EXERCISE DISCUSSION – CH5



5.1 A researcher obtains the following sample of genotypes from pandas living in two different geographic regions of China:

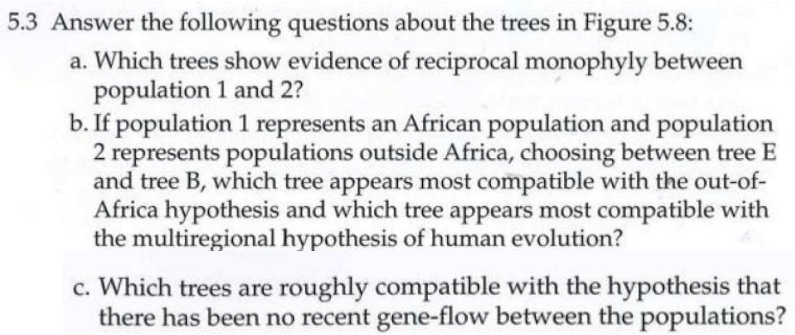
| *Genotype* | *AA* | *Aa* | *aa* |
|---|---|---|---|
| Region 1 | 12 | 22 | 6 |
| Region 2 | 32 | 6 | 2 |

Assume a simple divergence model in which the ancestral population size and the population sizes of both of the current populations are all equal to $2N = 10{,}000$. Furthermore, assume there has been no gene-flow since the time of divergence. Provide an estimate of the divergence time in number of years between the populations, assuming a generation time of one generation every five years.

5.2 Assuming an infinite sites model, mark the mutations in the sequences below on the tree as in Figure 5.2. If there are two or more possibilities for a mutation, mark all possible assignments on the tree. If the mutation in the site is not compatible with the tree under an infinite sites model, do not map the mutation on the tree.

Sequences

```
1  CCTCCAG
2  CGTCCTG
3  CCTGCAG
4  ACTGCAG
5  ACAGCTA
6  ACAGTGA
```

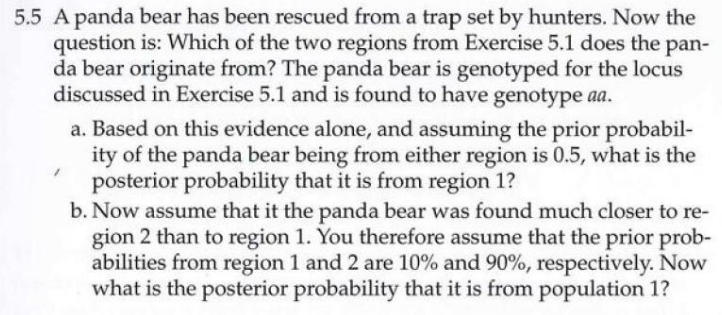


5.3 Answer the following questions about the trees in Figure 5.8:

a. Which trees show evidence of reciprocal monophyly between population 1 and 2?

b. If population 1 represents an African population and population 2 represents populations outside Africa, choosing between tree E and tree B, which tree appears most compatible with the out-of-Africa hypothesis and which tree appears most compatible with the multiregional hypothesis of human evolution?

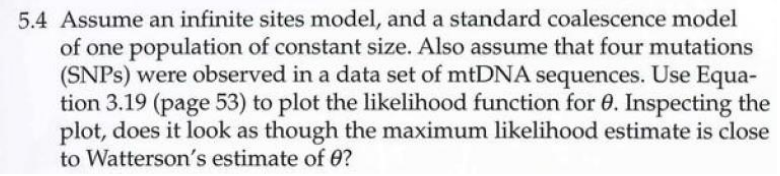
c. Which trees are roughly compatible with the hypothesis that there has been no recent gene-flow between the populations?



5.5 A panda bear has been rescued from a trap set by hunters. Now the question is: Which of the two regions from Exercise 5.1 does the panda bear originate from? The panda bear is genotyped for the locus discussed in Exercise 5.1 and is found to have genotype *aa*.

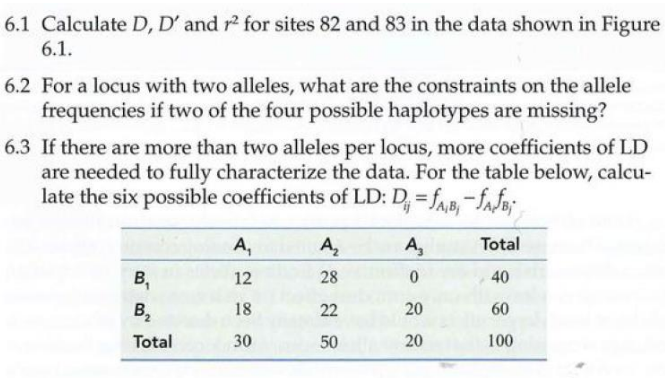
a. Based on this evidence alone, and assuming the prior probability of the panda bear being from either region is 0.5, what is the posterior probability that it is from region 1?

b. Now assume that it the panda bear was found much closer to region 2 than to region 1. You therefore assume that the prior probabilities from region 1 and 2 are 10% and 90%, respectively. Now what is the posterior probability that it is from population 1?



5.4 Assume an infinite sites model, and a standard coalescence model of one population of constant size. Also assume that four mutations (SNPs) were observed in a data set of mtDNA sequences. Use Equation 3.19 (page 53) to plot the likelihood function for $\theta$. Inspecting the plot, does it look as though the maximum likelihood estimate is close to Watterson's estimate of $\theta$?

# EXERCISE DISCUSSION – CH6

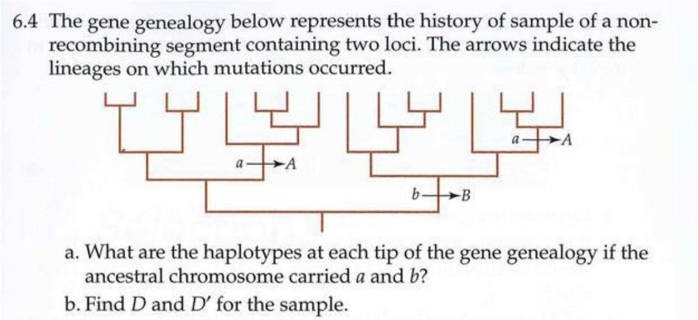


6.1 Calculate $D$, $D'$ and $r^2$ for sites 82 and 83 in the data shown in Figure 6.1.

6.2 For a locus with two alleles, what are the constraints on the allele frequencies if two of the four possible haplotypes are missing?

6.3 If there are more than two alleles per locus, more coefficients of LD are needed to fully characterize the data. For the table below, calculate the six possible coefficients of LD: $D_{ij} = f_{A_iB_j} - f_{A_i}f_{B_j}$.

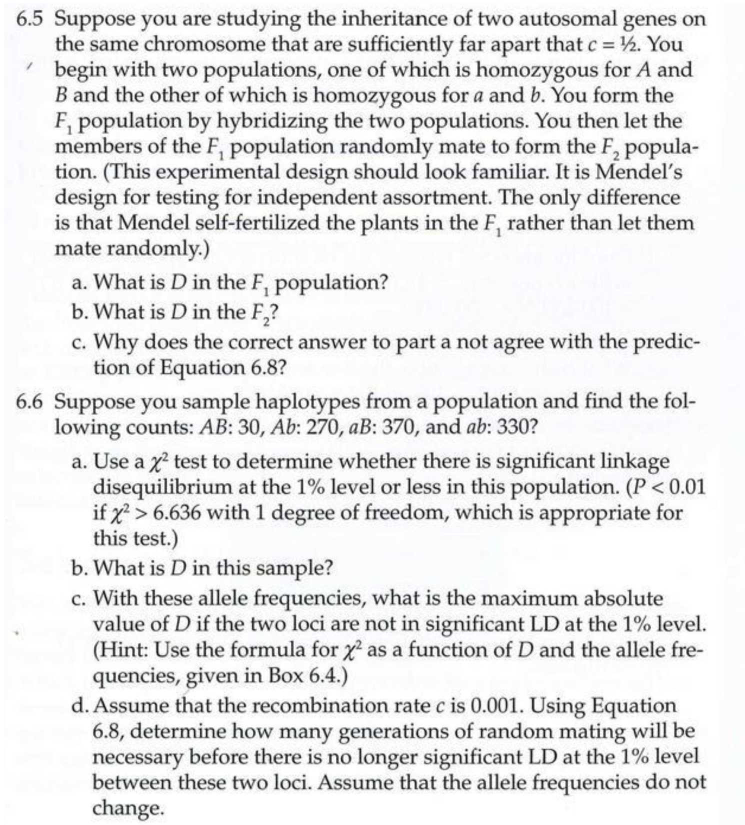
| | $A_1$ | $A_2$ | $A_3$ | Total |
|---|---|---|---|---|
| $B_1$ | 12 | 28 | 0 | 40 |
| $B_2$ | 18 | 22 | 20 | 60 |
| Total | 30 | 50 | 20 | 100 |



6.4 The gene genealogy below represents the history of sample of a nonrecombining segment containing two loci. The arrows indicate the lineages on which mutations occurred.

a. What are the haplotypes at each tip of the gene genealogy if the ancestral chromosome carried $a$ and $b$?

b. Find $D$ and $D'$ for the sample.



6.5 Suppose you are studying the inheritance of two autosomal genes on the same chromosome that are sufficiently far apart that $c$ = ½. You begin with two populations, one of which is homozygous for $A$ and $B$ and the other of which is homozygous for $a$ and $b$. You form the $F_1$ population by hybridizing the two populations. You then let the members of the $F_1$ population randomly mate to form the $F_2$ population. (This experimental design should look familiar. It is Mendel's design for testing for independent assortment. The only difference is that Mendel self-fertilized the plants in the $F_1$ rather than let them mate randomly.)

a. What is $D$ in the $F_1$ population?

b. What is $D$ in the $F_2$?

c. Why does the correct answer to part a not agree with the prediction of Equation 6.8?

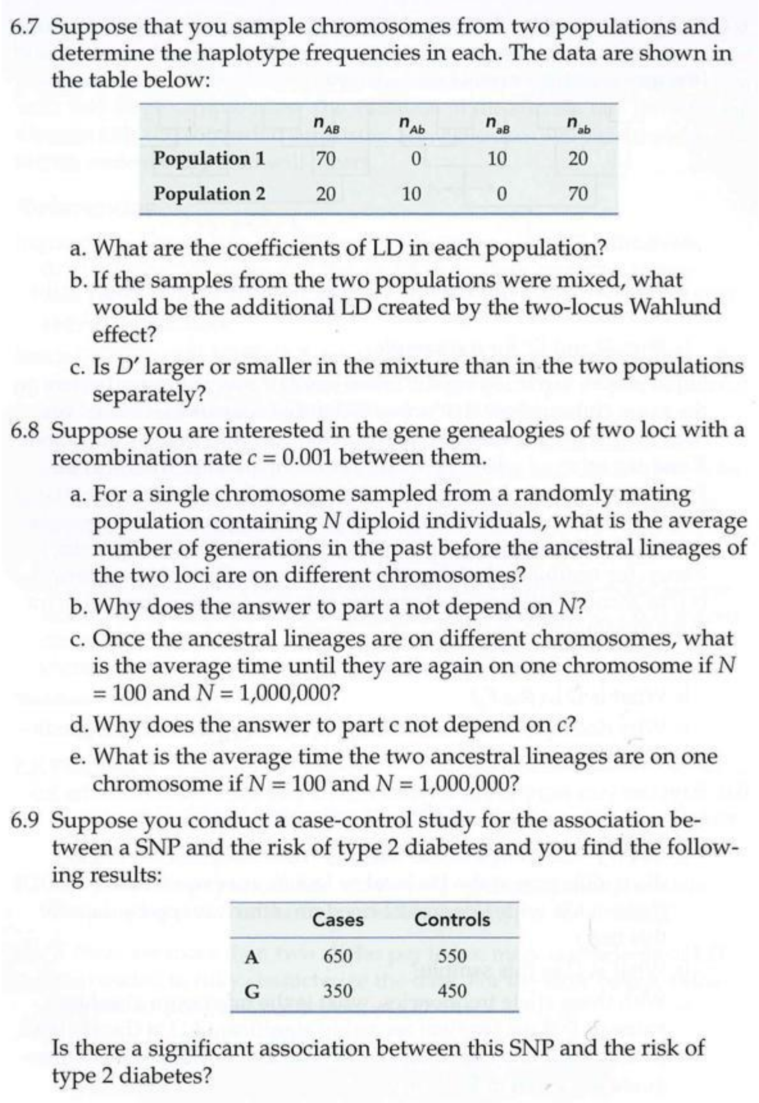
6.6 Suppose you sample haplotypes from a population and find the following counts: $AB$: 30, $Ab$: 270, $aB$: 370, and $ab$: 330?

a. Use a $\chi^2$ test to determine whether there is significant linkage disequilibrium at the 1% level or less in this population. ($P < 0.01$ if $\chi^2 > 6.636$ with 1 degree of freedom, which is appropriate for this test.)

b. What is $D$ in this sample?

c. With these allele frequencies, what is the maximum absolute value of $D$ if the two loci are not in significant LD at the 1% level. (Hint: Use the formula for $\chi^2$ as a function of $D$ and the allele frequencies, given in Box 6.4.)

d. Assume that the recombination rate $c$ is 0.001. Using Equation 6.8, determine how many generations of random mating will be necessary before there is no longer significant LD at the 1% level between these two loci. Assume that the allele frequencies do not change.



6.7 Suppose that you sample chromosomes from two populations and determine the haplotype frequencies in each. The data are shown in the table below:

| | $n_{AB}$ | $n_{Ab}$ | $n_{aB}$ | $n_{ab}$ |
|---|---|---|---|---|
| Population 1 | 70 | 0 | 10 | 20 |
| Population 2 | 20 | 10 | 0 | 70 |

a. What are the coefficients of LD in each population?

b. If the samples from the two populations were mixed, what would be the additional LD created by the two-locus Wahlund effect?

c. Is $D'$ larger or smaller in the mixture than in the two populations separately?

6.8 Suppose you are interested in the gene genealogies of two loci with a recombination rate $c = 0.001$ between them.

a. For a single chromosome sampled from a randomly mating population containing $N$ diploid individuals, what is the average number of generations in the past before the ancestral lineages of the two loci are on different chromosomes?

b. Why does the answer to part a not depend on $N$?

c. Once the ancestral lineages are on different chromosomes, what is the average time until they are again on one chromosome if $N$ = 100 and $N$ = 1,000,000?

d. Why does the answer to part c not depend on $c$?

e. What is the average time the two ancestral lineages are on one chromosome if $N$ = 100 and $N$ = 1,000,000?

6.9 Suppose you conduct a case-control study for the association between a SNP and the risk of type 2 diabetes and you find the following results:

| | Cases | Controls |
|---|---|---|
| A | 650 | 550 |
| G | 350 | 450 |

Is there a significant association between this SNP and the risk of type 2 diabetes?

# NEXT TIME

- More exercises!

- Menti quiz for this week – replacing recap (time-wise)

Mentimeter

- Weekly feedback :]

Mentimeter

AARHUS
UNIVERSITY