# Homework Excercises for Human Evolutionary Genetics

Alan R. Rogers[1] Jon Seger[2]

February 21, 2024

[1]Department of Anthropology, University of Utah, Salt Lake City, UT 84112
[2]Department of Biology, University of Utah, Salt Lake City, UT 84112

# Contents

# Introduction

This document contains homework assignments for *Human Evolutionary Genetics* (Anth/Biol 5221, at the University of Utah). The due dates of the assignments are given on the course syllabus. In some cases, more than one assignment may be due on the same day.

Answers to even-numbered problems are at the back of this document, beginning on p. 34.

## Notation for logarithms

Logarithms appear in various places in the following homework assignments. Usually, we are interested in the "natural" logarithm—the logarithm to base $e$, where $e \approx 2.72$. Our notation is as follows:

| Base | Notation |
|---|---|
| $e$ | $\log_e x$ or $\ln x$ or $\log x$ |
| 10 | $\log_{10} x$ |
| 2 | $\log_2 x$ |

Note that for us, $\log x$ means $\log_e x$. Some disciplines and some calculators use a different convention, in which $\log x$ means $\log_{10} x$, so be sure to understand what your own calculator does.

Our convention is also used within the Python programming language. For example,

```
>>> from math import *
>>> log(e**3)
3.0
>>> log10(10**2)
2.0
```

In this example, Python's `log` function inverts (reverses the effect of) raising $e$ to a power. This implies that Python's `log(x)` means $\log_e x$. Similarly, Python's `log10` function inverts the operation of raising 10 to a power, so `log10(x)` means $\log_{10} x$.

# Homework 1

# Probability

Each exercise is worth 10 points.

**Exercise 1.1.** *You toss a fair coin twice. What is the probability of observing two heads?*

**Exercise 1.2.** *With the same coin experiment, what is the probability of a head and a tail (in that order)?*

**Exercise 1.3.** *What is the probability of a tail and then a head?*

**Exercise 1.4.** *What about a head and a tail in either order?*

**Exercise 1.5.** *You toss a coin three times and observe 2 heads and a tail in some order. In how many ways can this happen? (In other words, how many sequences like "HHT" contains 2 "H"s and one "T"?)*

**Exercise 1.6.** *You toss a coin four times and observe 2 heads and 2 tails in some order. In how many ways can this happen? (In other words, how many sequences like "HHTT" contains 2 "H"s and 2 "T"s?)*

**Exercise 1.7.** *What is the probability of observing 2 heads and a tail (order unspecified) in 3 tosses.*

**Exercise 1.8.** *What is the probability of observing 2 heads and 2 tails (order unspecified) in 4 tosses.*

**Exercise 1.9.** *You toss two fair dice, one red and one black. What is the probability that you observe either a red 4 or a black 6 (or both)?*

**Exercise 1.10.** *You toss one fair die. What is the probability that you observe either a 4 or a 6?*

**Exercise 1.11.** *You toss two fair dice, one red and one black. What is the probability that you observe both a red 4 and a black 6?*

Imagine a modified version of Kerrich's urn experiment in which each trial begins with 3 balls of each color (red and black).

**Exercise 1.12.** *What is the probability that, in a single trial, both of the balls drawn are red?*

**Exercise 1.13.** *What is the probability that, in a single trial, the first ball is red and the second black?*

You toss a fair coin 3 times. You receive \$1 for each head and nothing for tails. Let $X$ represent the number of dollars you receive.

**Exercise 1.14.** *For this random variable, what are the possible values and the probability of each value? In other words, what is the probability distribution of $X$?*

**Exercise 1.15.** *What is the mean?*

**Exercise 1.16.** *What is the variance?*

In JEPr, you saw several tables of counts. Here is another:

| | My mood | | |
|---|---|---|---|
| Weather | Happy | Sad | Sum |
| Rain | 30 | 70 | 100 |
| Sun | 90 | 10 | 100 |
| Sum | 120 | 80 | 200 |

In this table, each cell counts the number of days during which (a) it rained and I was happy, (b) it rained and I was sad, and so on. The table tells us that it rained on 100 days, was sunny on another 100, and that I was happy 90% of the time on the sunny days but only 30% of the time on rainy ones. Use this table in the following exercises.

**Exercise 1.17.** *What is the* unconditional *relative frequency of sad?*

**Exercise 1.18.** *What is the* unconditional *relative frequency of rainy?*

**Exercise 1.19.** *What is the* conditional *relative frequency of sad given rainy?*

**Exercise 1.20.** *What is the* conditional *relative frequency of rainy given sad?*

It should turn out that $f(\text{rain}|\text{sad}) \neq f(\text{rain})$. In a statistical sense, therefore, the weather depends on my mood. Yet this does *not* imply that my mood has any causal effect. Statistical dependence should not be confused with causation.

**Exercise 1.21.** *Consider a locus with two alleles, A and a, whose frequencies are $p$ and $q = 1-p$. Suppose that you draw one gene copy from this population at random. Let $X = 1$ if that gene is a copy of A and let $X = 0$ otherwise. What are (a) the mean and (b) the variance of $X$, as a function of $p$?*

**Exercise 1.22.** *Suppose we form diploid offspring by sampling with replacement from the parental gene pool. Then in any random offspring, the number, $Y$, of copies of allele A is the sum of two independent random variables, each of which equals 1 with probability $p$ and 0 with probability $1 - p$. What does this imply about the mean and variance of $Y$?*

Consider a particular nucleotide site in the DNA of some hypothetical population. In this population, 80% of chromosomes have the nucleotide adenine (A) while 20% have guanine (G). In other words, the relative frequency of allele A is 0.80. We take samples from this population, each of which consists of two chromosomes drawn independently and at random. In each sample, we observe either AA, AG, or GG. The number of A's in this sample is a random variable, which I will call $X$. This variable can take values 0, 1, or 2.

**Exercise 1.23.** *What is the distribution of $X$?*

**Exercise 1.24.** *What is the mean?*

**Exercise 1.25.** *What are the variance and standard deviation?*

**Exercise 1.26.** *Sketch a histogram of the distribution of $X$.*

JEPr describes Bortkiewicz's data on the frequency of deaths caused by mule kicks in the Prussian army. As you will recall, those data are approximately Poisson, and the mean number of deaths per corps-year was $\lambda = 0.61$. Use the Poisson distribution formula to answer the following questions.

**Exercise 1.27.** *What is the probability that there will be zero such deaths during a given year for a given corps?*

**Exercise 1.28.** *What is the probability that there will be one such death during a given year for a given corps?*

**Exercise 1.29.** *What is the probability that there will be* AT LEAST *one such death during a given year for a given corps?*

**Exercise 1.30.** *What is the probability that there will be 2 such deaths during a given year for a given corps?*

Suppose that the number of offspring per female is Poisson, and that the average female has two offspring. (This keeps the population from growing or shrinking.) We are assuming, in other words, that the Poisson distribution has mean $\lambda = 2$.

**Exercise 1.31.** *What is the probability that a random female will have no children at all?*

**Exercise 1.32.** *What is the probability that a random female will have one child?*

**Exercise 1.33.** *What is the probability that a random female will have* AT LEAST *one child?*

In Kerrich's urn experiment, suppose you get \$1 for each red ball and \$0 for each green one, and let $X$ and $Y$ represent the dollars you receive on the two draws within a single trial of the experiment.

**Exercise 1.34.** *Write down the probability distribution of $X$ and $Y$ in tabular form. Your table should have columns for $X$, for $Y$, and for the joint probability of $X$ and $Y$, i.e.* $\Pr[X, Y]$.

**Exercise 1.35.** *What are the expected values of $X$ and of $Y$?*

**Exercise 1.36.** *What are the variances of $X$ and of $Y$?*

**Exercise 1.37.** *What is the covariance of $X$ and $Y$?*

Imagine an urn with $N$ balls, of which 1 is red and the rest are green. You draw 2 balls from the urn at random *without* replacement. Let $X = 1$ if the first ball is red and $X = 0$ otherwise. Define $Y$ similarly for the second ball.

**Exercise 1.38.** *Draw a tree to represent the probabilities in this experiment. Use Fig. 2 of JEPr as a model.*

**Exercise 1.39.** *What are the mean and variance of $X$?*

**Exercise 1.40.** *What are the mean and variance of $Y$?*

**Exercise 1.41.** *What is the covariance of $X$ and $Y$? (Hint: the previous two questions used a table to represent the probability distribution of $(X, Y)$. Add a column to this table to represent the product, $XY$.)*

Suppose that, in a class of 50 students, 20 are women.

**Exercise 1.42.** *If we choose a student at random from the class, what is the probability that this student is a woman?*

**Exercise 1.43.** *If we choose 2 students from this class at random* without *replacement, what is the probability that both are women?*

JEPr discussed the following probability distributions: (1) binomial, (2) Bernoulli, (3) Poisson, (4) uniform, (5) exponential, and (6) normal. Which of these choices would be most appropriate in each of the following contexts? (Just write down the name of the appropriate distribution.)

**Exercise 1.44.** *The weight of a mouse, selected at random from those that live in this building.*

**Exercise 1.45.** *The number of neutral mutations on a gene genealogy of known length.*

**Exercise 1.46.** *The number of copies of $A_1$ (a neutral allele) on some small island in the South Pacific, assuming that we know the size of this population and the allele frequency among the parents.*

# Homework 2

# Random mating

Table 2.1: Transferrin genotype frequencies in a baboon troop [4, p. 56].

| | Number of | | |
|---|---|---|---|
| Genotype | individuals | allele $C$ | allele $D$ |
| $CC$ | 80 | 160 | 0 |
| $CD$ | 15 | 15 | 15 |
| $DD$ | 5 | 0 | 10 |
| Total | 100 | 175 | 25 |

## 2.1 Frequencies of alleles and genotypes

*Transferrin* is a protein involved in iron transport. Table 2.1 shows the number of individuals in a baboon troop, grouped by transferrin genotype. The relative frequency of each genotype is simply the number of individuals of that genotype divided by the total number of individuals.

**Exercise 2.1.** *What is the relative frequency, $P_{CC}$, of the $CC$ genotype?*

**Exercise 2.2.** *What is the relative frequency of the $CD$ genotype?*

**Exercise 2.3.** *What is the relative frequency of the $DD$ genotype?*

There are two ways to calculate the allele frequency within a sample:

1. Divide the number of copies of one allele by the total number of gene copies in the sample.

2. If the locus has just two alleles, we can use the formula
$$p_1 = P_{11} + P_{12}/2 \tag{2.1}$$
where $P_{11}$ and $P_{12}$ are the frequencies of genotypes $A_1A_1$ and $A_1A_2$, and $p_1$ is the frequency of allele $A_1$.

**Exercise 2.4.** *Use both methods to calculate the frequency, $p_C$ of the $C$ allele.*

**Exercise 2.5.** *Use both methods to calculate the frequency, $p_D$ of the $D$ allele.*

**Exercise 2.6.** *Prove that the two methods are equivalent.*

At some SNP locus, a sample of 100 individuals includes 15 copies of genotype "CC," 50 of "CT," and 35 of "TT." Please use these data in answering the following:

**Exercise 2.7.** *What is the number of gene copies in this sample?*

**Exercise 2.8.** *What are the allele frequencies $p$ and $q$ of the two nucleotides (C and T)?*

**Exercise 2.9.** *What is the expected heterozygosity under random mating?*

**Exercise 2.10.** *What is the observed heterozygosity?*

## 2.2 Using $F$-statistics

The observed and HW genotype frequencies are

| | Genotype Frequencies | |
|---|---|---|
| Genotype | Observed | Hardy-Weinberg |
| $A_1A_1$ | $P_{11}$ | $p^2$ |
| $A_1A_2$ | $P_{12}$ | $2pq$ |
| $A_2A_2$ | $P_{22}$ | $q^2$ |

In the HW formulas, $p$ and $q$ are the allele frequencies of the *population* and are unknown. We can, however, estimate them from our data by $\hat{p} = P_{11} + P_{12}/2$ and $\hat{q} = 1 - \hat{p}$. Plugging these estimates into the HW formulas will provide estimates of the HW genotype frequencies, which can then be compared with $P_{11}$, $P_{12}$, and $P_{22}$.

To measure the deviation between observed and expected genotype frequencies, it is convenient to define a variable, $F$, which satisfies

$$\begin{aligned} P_{11} &= p^2 + pqF \\ P_{12} &= 2pq - 2pqF \\ P_{22} &= q^2 + pqF \end{aligned}$$

The three equations above provide three different ways to estimate $F$:

$$F = \frac{P_{11} - p^2}{pq} \tag{2.2}$$

$$F = -\frac{P_{12} - 2pq}{2pq} \tag{2.3}$$

$$F = \frac{P_{22} - q^2}{pq} \tag{2.4}$$

All three formulas produce the same number.

Consider the following data

| Genotype | Number of copies |
|---|---|
| $A_1A_1$ | $n_{11} = 100$ |
| $A_1A_2$ | $n_{12} = 100$ |
| $A_2A_2$ | $n_{22} = 100$ |

In other words, there are 100 copies of each of the three genotypes.

**Exercise 2.11.** *Use these data to calculate both the observed and the expected genotype frequencies.*

**Exercise 2.12.** *Next, calculate $F$ using each of formulas 2.2, 2.3, and 2.4.*

**Exercise 2.13.** *In a sample of 72 individuals, we have the following genotype counts: $n_{11} = 0$, $n_{12} = 5$, and $n_{22} = 67$. What are $P_{11}$, $P_{12}$, $P_{22}$, $p_1$, and $F$?*

**Exercise 2.14.** *In another sample, we have the following genotype counts: $n_{11} = 53$, $n_{12} = 19$, and $n_{22} = 1$. What are $P_{11}$, $P_{12}$, $P_{22}$, $p_1$, and $F$?*

## 2.3 Final words

We have glossed over various technicalities in this homework assignment. For example, the HW formulas $p^2$, $2p(1-p)$, and $(1-p)^2$ are correct when $p$ is the allele frequency *in the population*, but we have encouraged you to substitute the estimate of $p$ obtained from the data. This introduces a small bias, which can be corrected by using suitably modified versions of the HW formulas. For details, see Morton et al. [6, p. 509]. Unless your sample is very small, however, the effect of this correction is small.

# Homework 3

# Genetic Drift

Each question is worth 10 points.

## 3.1 Drift acting alone

In the first few problems, we assume that the only force at work is genetic drift. There is no selection, no mutation, and no migration. The first few problems all involve the equation

$$H_t = H_0(1 - 1/2N)^t \tag{3.1}$$

which describes the decay of heterozygosity in a randomly-mating population.

**Exercise 3.1.** *In the urn model of genetic drift, there are $2N$ balls in the urn of which a fraction $p$ are black. The remaining fraction, $1-p$, are white. Suppose we draw $2N$ balls from the urn with replacement. Of the balls that we draw, a random number $X$ are black, and the remaining $2N - X$ are white. What are the* mean *and* variance *of $X$? (Your answers should be in terms of $2N$ and $p$.)*

**Exercise 3.2.** *Consider a very small population in which $2N = 10$, mating is at random, and genetic drift is the only evolutionary force at work. If heterozygosity equals $1/2$ in one generation, what is its expected value in the following generation?*

**Exercise 3.3.** *Consider a larger population in which $2N = 1000$, mating is at random, and genetic drift is the only evolutionary force at work. If heterozygosity equals $1/2$ in one generation, what is its expected value in the following generation?*

**Exercise 3.4.** *In Series I of Buri's experiment, the initial heterozygosity was $H_0 = 0.514$ and in the 18th generation it was $H_{18} = 0.183$. Take these values as given and solve equation 3.1 for $2N$. In your answer, include two digits to the right of the decimal point. How does your estimate compare to the one ($2N = 18$) that Buri got?*

**Exercise 3.5.** *Suppose that $2N = 18$ on average, but that the value was not constant. Specifically, suppose that $2N = 26$ for the first 9 generations and $2N = 10$ for the last 9. (Note that these numbers average to 18.) What value would you predict for $H_{18}$? Compare this to the answer you get when $2N = 18$ in every generation. Which answer is higher? How does variation in population size affect the decay of heterozygosity?*

**Exercise 3.6.** *Suppose now that the population size alternated between $2N = 26$ and $2N = 10$. It was 26 in the 1st generation, 10 in the 2nd, 26 again in the 3rd, and so on. What is the expected heterozygosity in generation 18? (Hint: Think carefully about equation 3.1. There is a way to answer this without doing any additional calculations.)*

**Exercise 3.7.** *In Buri's experiment, how many generations would it take before $H_t = 0.1$? (Assume $2N = 18$.)*

**Exercise 3.8.** *Now rearrange equation 3.1 so that $t$ is on the left side, and everything else is on the right. (Hint: The right side will be an algebraic expression involving the symbols $H_0$, $H_t$, and $2N$. There will be logarithms.) Use your answer to this question to check your answer to the preceding one.*

**Exercise 3.9.** *Suppose that a catastrophe reduces diploid population size from 1024 to 2. Thereafter, the population doubles every generation until it reaches its original size. What fraction of the original heterozygosity still remains? To do this precisely, you would want to ask whether the species had separate sexes. We haven't taught you how to do this, so please assume that these organisms have monoecious sexual reproduction and mate at random. Then the theory we have taught in class works all the way down to $N = 2$. Feel free to do this either by hand or in Python.*

## 3.2 Mutation and drift

Gillespie shows that, under the "infinite alleles" model of mutation, heterozygosity ($H$) evolves under the combined effects of mutation and genetic drift toward an equilbrium value,

$$\hat{H} = \frac{\theta}{1+\theta} \tag{3.2}$$

where $\theta = 4Nu$.

**Exercise 3.10.** *At autosomal loci, the mutation rate per gene is often around $u = 10^{-6}$ per generation. Let us suppose this is so for Buri's flies. Assuming once again that the effective population size is $2N_e = 18$, calculate the equilibrium heterozygosity in Buri's experiment. This value represents the expected heterozygosity within each bottle in the long run—after many generations of simulaneous mutation and drift. At this equilibrium, how many of the 16 flies in each bottle would be likely to be heterozygous? Round your answer to the nearest integer.*

**Exercise 3.11.** *Solve Eqn. 3.2 algebraically for $\theta$ as a function of $\hat{H}$.*

The model of infinite alleles, which underlies Eqn. 3.2, is plausible model for protein-coding loci, which may contain thousands of nucleotide sites. The number of possible alleles is so large that we get a good approximation by taking that number to be infinite. This approximation works poorly, however, for individual nucleotide sites. These have at most 4 alleles—A, T, G, and C. If there are $k$ possible alleles, and each allele mutates with equal probability to each of the other alleles, then Eqn. 3.2 becomes

$$\hat{H} = \frac{\theta}{1+\theta k/(k-1)} \tag{3.3}$$

This model is also unrealistic, however, because it is seldom true that each allele mutates with equal probability to each other allele. Typically, the most common type of mutation is the "transition," which toggles the nucleotide back and forth either between A and G or between C and T. In human mitochondrial DNA, for example, the vast majority of mutations are of this type. When this is so, we get a reasonable approximation by assuming that $k = 2$ in Eqn. 3.3. Let us call this the "symmetric biallelic" model of mutation, because it implies that each polymorphic site has 2 alleles, each of which mutates to the other at the same rate. For this model,

$$\hat{H} = \frac{\theta}{1+2\theta}$$

**Exercise 3.12.** *What is the largest value that $\hat{H}$ can take in the model of infinite sites?*

**Exercise 3.13.** *What is the largest value that $\hat{H}$ can take in the symmetric biallelic model?*

**Exercise 3.14.** *Suppose that random pairs of mitochondria differ on average at 0.5% of nucleotide sites, that the population is at mutation-drift equilibrium, that the mutation rate is $u = 2 \times 10^{-7}$ per nucleotide per generation, and that the symmetric biallelic model of mutation applies. What does this imply about the effective number of females? (Hint, we can use Eqn. 3.3, provided that we interpret the number, $2N$, of gene copies in the population as the effective number of females.)*

**Exercise 3.15.** *Suppose that the average individual is heterozygous at $1/3$ of protein-coding loci, that the population is at mutation-drift equilibrium, that the mutation rate is $u = 10^{-6}$ per locus per generation, and that the infinite alleles model of mutation applies. What does this imply about $N$?*

# Homework 4

# Gene Genealogies

In the following exercises, assume that the population size is constant and that the genetic variation under study is selectively neutral. Each exercise is worth 10 points.

The exercises below are all based on material in *Lecture Notes on Gene Genealogies* (LNGG). Those in section 4.1 are based on chapter 4 of LNGG, whereas those in section 4.2 are based on chapter 5. There is one exception to this: exercise 4.5 is based on chapter 5.

## 4.1 Gene genealogies without mutation

**Exercise 4.1.** *Figure 4.1 (on page 11) shows a made-up gene genealogy of 6 DNA sequences. The branches are labeled with capital letters. (a) Which sequences would carry the derived allele if a mutation occurred on branch H? (b) On branch B? (c) On branch G?*

For the exercises in section 4.1, assume that $2N = 5000$.

**Exercise 4.2.** *What is the expected duration in generations of coalescent intervals with (a) 2 lines of descent, (b) 10 lines of descent, and (c) 1000 lines of descent?*

**Exercise 4.3.** *In an interval with 2 lines of descent, what are the* mean*, the* variance *and the* standard deviation *of the interval's duration? Hint: the duration of the interval is an exponentially-distributed random variable. This distribution is discussed in JEPr.*

For exercises 4.4 and 4.5, assume that you have a sample consisting of $K = 3$ gene copies chosen at random from a population of size $2N = 5000$.

**Exercise 4.4.** *What is the expected depth of the gene genealogy? In other words, the expected age in generations of the last common ancestor?*

**Exercise 4.5.** *What is the expected total branch length the gene genealogy? In other words, the expected sum of the lengths of all branches throughout the genealogy? (This problem is based on material in section 5.1 of* Lecture Notes on Gene Genealogies*.)*

## 4.2 Gene genealogies with mutation

**Exercise 4.6.** *Suppose that $\theta = 10$, indicating either that the population is very large or the mutation rate is high. How many mutations should occur on average in coalescent intervals with (a) 2 lines of descent, (b) 10 lines of descent, and (c) 1000 lines of descent?*

**Exercise 4.7.** *(a) Draw a gene tree with 4 tips. What are (b) the expected lengths of each interval (assuming the population consists of $N$ diploid individuals), (c) the expected depth of the tree, and (d) the expected total tree length (i.e. the sum of all branch lengths)? (Hint: your answers will be in terms of the unknown quantity $N$.)*

**Exercise 4.8.** *How many mutations would you expect to see on such a tree, assuming a mutation rate of $u = 1/1000$ and a haploid population size of $2N = 5000$?*

**Exercise 4.9.** *Now suppose that you doubled the sample size from 4 to 8. Don't draw the tree. Just calculate the expected number of mutations. How much did the number of mutations increase?*

**Exercise 4.10.** *What is the ratio between the expected value of $S$ in a sample of 10 DNA sequences and the expected value in a sample of 20?*

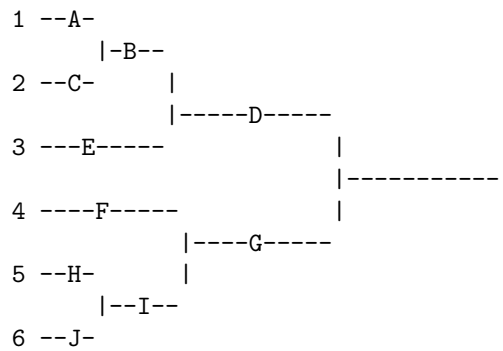



Figure 4.1: A genealogy with labeled branches. The samples are numbered (1–6) and the branches are labeled (A–J).

```
seq01 AATATGGCAC CTCCCAACCC TCTAGCATAT ACCACTTACA
seq02 .......T.. .C......TG C......C.. ..........
seq03 ..C....... .......... .......... ..........
seq04 .......T.. .C......TG C......... G.........
seq05 .......... .......... .......... ..........
seq06 .....A.... ........T. C......... G....C....
seq07 ..C....T.. .C......TG C......... G.........
seq08 .....A.T.. TC......TG C......... G.........
seq09 .......... .......... C......... ..........
seq10 .G...A.... ........T. C......C.. .T....C..G
```

Table 4.1: Data set A. Periods indicate sites that are identical to `seq01`.

```
seq01 TGCCACTCCA ATCTCTCGCC AGATGGCATG CCTTATCGCG
seq02 .......... G......... .A.C...GCA T.........
seq03 .......... G......A.. .A.C...GC. T....C....
seq04 C..TG..T.. .C.....A.. G......C.. TT.C......
seq05 CA.TG..T.. .C....TA.. G......CC. TT.C......
seq06 C...G..... CC..T..A.A ....AA.C.. TT..G.....
seq07 CA.TG..... CC..TC.A.A ...CA..CC. TT...C....
seq08 CA.TG...T. GCT....A.. G..C..TC.. T.......T.
seq09 CA.TG..... GC.C..T... ...CA..C.A T.........
seq10 CA.TG...T. .C.....A.. G..C...C.. T....C....
```

Table 4.2: Data set B

**Exercise 4.11.** *Use data set A to calculate $\pi$, the mean number of nucleotide site differences per sequence* and *per site between pairs of sequences.*

In calculating $\pi$ (the number of pairwise differences), data set B poses a new problem: two of the sites (11 and 27) have more than two nucleotides. This will not cause a problem if you do the calculation the laborious way, by counting the differences between each pair of sequences. But if you use the easier site-by-site method described in the text, you need to know how to deal with such sites.

Consider site 11, which has 4 *A*s, 4 *G*s, and 2 *C*s. We are interested only in the pairs that have different nucleotides. In other words, we are only interested in pairs of type *AG*, *AC*, or *GC*. The number of *AG* pairs is $4 \times 4 = 16$. This follows because each of the 4 As can combine with each of the 4 *G*s. Similarly the number of *AC* pairs is $4 \times 2 = 8$, and the number of *GC* pairs is $4 \times 2 = 8$. This site therefore contributes $16 + 8 + 8 = 32$ to our count of differences.

At site 27, we have 1 *A*, 2 *G*s, and 7 *C*s. This gives $1 \times 2 = 2$ *AG*s, $1 \times 7 = 7$ *AC*s, and $2 \times 7 = 14$ *GC*s, so the total contribution from site 27 is $2 + 7 + 14 = 23$.

**Exercise 4.12.** *Use data set B to calculate $\pi$, the mean number of nucleotide site differences per sequence* and *per site between pairs of sequences.*

**Exercise 4.13.** *Calculate the number $S$ of segregating sites from data set A.*

**Exercise 4.14.** *Calculate the number $S$ of segregating sites from data set B.*

**Exercise 4.15.** *Estimate $\theta$ (per sequence and per site) from the value of $\pi$ you got from data set A.*

Table 4.3: Values of $\sum_{k=1}^{K-1} 1/k$ for various values of $K$

| $K$ | $\sum_{k=1}^{K-1} 1/k$ | $K$ | $\sum_{k=1}^{K-1} 1/k$ |
|---|---|---|---|
| 2 | 1.0000 | 12 | 3.0199 |
| 3 | 1.5000 | 13 | 3.1032 |
| 4 | 1.8333 | 14 | 3.1801 |
| 5 | 2.0833 | 15 | 3.2516 |
| 6 | 2.2833 | 16 | 3.3182 |
| 7 | 2.4500 | 17 | 3.3807 |
| 8 | 2.5929 | 18 | 3.4396 |
| 9 | 2.7179 | 19 | 3.4951 |
| 10 | 2.8290 | 20 | 3.5477 |
| 11 | 2.9290 | 21 | 3.5977 |

**Exercise 4.16.** *Estimate $\theta$ (per sequence and per site) from the value of $\pi$ you got from data set B.*

**Exercise 4.17.** *Estimate $\theta$ from the value of $S$ you got from data set A. (Hint: Use table 4.3.)*

**Exercise 4.18.** *Estimate $\theta$ from the value of $S$ you got from data set B.*

**Exercise 4.19.** *What can you infer from the similarity (or the dissimilarity) of the two estimates of $\theta$ you got from data set A. (Hint: Remember that it is only reasonable to compare values in the same units. Don't compare differences per site with differences per sequence.)*

**Exercise 4.20.** *What can you infer from the similarity (or the dissimilarity) of the two estimates of $\theta$ you got from data set B.*

**Exercise 4.21.** *Calculate the folded site frequency spectrum for data set A.*

**Exercise 4.22.** *Calculate the folded site frequency spectrum for data set B.*

**Exercise 4.23.** *Use the $\hat{\theta}_S$ for data set A to calculate the theoretical folded spectrum for neutral loci in a population of constant size. (You will want the value of $\hat{\theta}_S$* per sequence*, not per site.) How well do the observed and theoretical spectra match?*

**Exercise 4.24.** *Use the $\hat{\theta}_S$ for data set B to calculate the theoretical folded spectrum. (You will want the value of $\hat{\theta}_S$* per sequence*, not per site.) How well do the observed and theoretical spectra match?*

**Exercise 4.25.** *Sketch a gene genealogy for a sample of 10 sequences taken from a population that grew suddenly from a small size to a very large size 6 units of mutational time ago.*

**Exercise 4.26.** *Looking back from the present to the time of the population expansion, the number of new mutations on a single line of descent will be 3 on average, but not in every case. What probability distribution would best describe the variation in the number of mutations among independent lines of descent?*

**Exercise 4.27.** *Sketch the distribution of pairwise differences (i.e. the mismatch distribution) that you would expect to see in this sample. Assume that the infinite sites model is a good approximation here. In other words, don't worry about multiple hits. Be sure to label the X axis in a way that indicates the mode of the distribution.*

**Exercise 4.28.** *In sequence data from the control region of human mitochondrial DNA, gene diversity ($\pi$) is often around 0.01 for European and Asian samples and around 0.02 for African ones. What does this imply about the numerical values of $\theta = 4Nu$ in Europe and Africa?*

The parameter $\theta$ can be estimated either from

$$\hat{\theta}_\pi = \pi \qquad \text{or from} \qquad \hat{\theta}_S = S \Big/ \sum_{k=1}^{K-1} 1/k,$$

where $\pi$ is the mean pairwise difference per sequence, $S$ is the number of segregating sites and $K$ is the number of gene copies in the sample. In a population of constant size, these statistics are often similar in size. In a population whose size has changed, they can be dramatically different.

The reason involves the effect of population growth on the site frequency spectrum. As explained in section 7.7 of Rogers [9], we see an excess of singleton sites in populations that have expanded in size at some time since the last common ancestor of the gene genealogy. To understand how population growth affects $\hat{\theta}_\pi$ and $\hat{\theta}_S$, we need to think about how these statistics are affected by singleton sites.

**Exercise 4.29.** *What does a singleton site contribute to the value of $\hat{\theta}_\pi$ in a sample of $K$ gene copies?*

**Exercise 4.30.** *What does a singleton site contribute to the value of $\hat{\theta}_S$ in a sample of $K$ gene copies?*

**Exercise 4.31.** *Make a table with three columns: the first for the number, $K$ of gene copies*

*in the sample, the second for the effect of a singleton site on $\hat{\theta}_S$ and the third for the effect on $\hat{\theta}_\pi$. Consider $K$ values of 2, 3, 4, 10, and 100. For what values of $K$ does a singleton site have a larger effect on $\hat{\theta}_S$ than on $\hat{\theta}_\pi$?*

**Exercise 4.32.** *In a population that has expanded dramatically in size, it is usually true that $\hat{\theta}_S > \hat{\theta}_\pi$. Why?*

# Homework 5

# Mismatch Distribution and Spectrum

Each exercise is worth 10 points.

**Exercise 5.1.** *Describe in words the expected mismatch distribution of a population that experienced a major episode of growth 8 units of mutational time ago, but was otherwise constant in size. (Hint: a pair of gene copies that has been separated for 8 units of mutational time will on average differ by 8 mutations.)*

**Exercise 5.2.** *How should the unfolded site frequency spectrum of this expanded population differ from that of a population with a long history of constant size?*

Here is a set of 10 DNA sequences, each with 10 sites:

```
seq0      AAAATAAAAA
seq1      ....A....T
seq2      TTT.A...T.
seq3      .TT.AT...T
seq4      ...TA...T.
seq5      ...T.T.TTT
seq6      TT..A.....
seq7      ....A.....
seq8      .T..A.....
seq9      ......T.T.
```

**Exercise 5.3.** *Use these data to calculate a mismatch distribution. (Remember that the first entry of the mismatch distribution is the number of pairs of sequences that differ at 0 nucleotide sites.)*

**Exercise 5.4.** *To check your mismatch distribution, make sure that the sum of the counts in the mismatch distribution is equal to the number of pairs of individuals in the data, i.e. to $K(K-1)/2$, where $K$ is the number of sequences.*

**Exercise 5.5.** *As a second check, calculate $\pi$, the mean pairwise difference, from the mismatch distribution as follows:*

$$\pi = \left(\frac{K(K-1)}{2}\right)^{-1} \sum_i iF_i$$

*where $F_i$ is the number of pairs of sequences that differ by $i$ nucleotide sites and $K$ is the number of sequences.*

**Exercise 5.6.** *Calculate $\pi$ once again using the easy method described in my chapter on Descriptive Statistics for DNA Sequences. Make sure that the two ways of calculating $\pi$ give the same answer.*

Tables A and B contain two additional sets of data, each with 20 sequences and 50 sites. For your convenience, we have also given you the mismatch distribution of each dataset. The first entry in the mismatch distribution is the number of pairs of sequences in the data that differ by 0 sites; the second entry counts the pairs that differ by 1 site, and so on.

These data sets were generated by computer simulations. For one data set, the simulation assumed a history of constant population size. For the other, the simulation assumed that the population experienced an episode of sudden growth but was constant in size before and after this episode. For the expanded population, the population history parameters were: $\theta_0 = 1$, $\theta_1 = 1000$, and $\tau = 5$. For the stationary population, there is only one population history parameter: $\theta = 5.98$.

Your job is to figure out which data set came from which population.

**Exercise 5.7.** *Plot the mismatch distribution of data set A.* ***Do not calculate the distribution***

***yourself.*** *The numbers are provided for you below the data set.*

**Exercise 5.8.** *Repeat exercise 5.7 using data set B.*

**Exercise 5.9.** *Calculate the folded observed site-frequency spectrum of data set A.*

**Exercise 5.10.** *Calculate the folded observed site-frequency spectrum of data set B.*

The next few exercises will ask you to calculate the folded spectrum that is expected under the hypothesis of neutrality and constant population size—the "theoretical" folded spectrum. Here's how to calculate the expected values: Let $s_i$ represent the $i$th entry of the empirical spectrum in a sample of $K$ DNA sequences. Its expected value is

$$E[s_i] = \theta/i + \theta/(K-i)$$

for positive integers $i$ that are less than $K/2$.

If the number of sequences is even, then the spectrum has an additional term at which $i = K/2$. For that term,

$$E[s_i] = \theta/i$$

This issue arises with the present data sets, both of which have an even number of sequences.

To calculate numerical values, replace $\theta$ in these expressions with its estimate,

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{K-1} 1/i}$$

where $S$ is the number of segregating sites. In our two data sets, $K = 20$, so the sum in the denominator is $\sum_{i=1}^{19} 1/i = 3.5477$. We mention this merely to save you the trouble of calculating it. With Python, however, the calculation is easy: `sum([1/i for i in range(1,20)])`.

**Exercise 5.11.** *Calculate the folded theoretical site-frequency spectrum of data set A—the spectrum expected under selective neutrality and constant population size.*

**Exercise 5.12.** *Calculate the folded theoretical site-frequency spectrum of data set B—the spectrum expected under selective neutrality and constant population size.*

**Exercise 5.13.** *Plot the observed and theoretical spectra for data set A.*

**Exercise 5.14.** *Plot the observed and theoretical spectra for data set B.*

**Exercise 5.15.** *Use these results to figure out which data set came from an expanded population and which came from a stationary population.*

Table 5.1: Polymorphic Sites from Data Set A

```
%         0123456789 0123456789 0123456789 0
seq0      TAAAAATATA ATAAAAAATA AAAAAAAAAA A
seq1      A.T.......  ........A.  ......T... .
seq2      A...T.....  ...T....A.  ......T... .
seq3      A.........  .....T.TA.  ......T... .
seq4      A.....AT..  ........A.  ......TTT. .
seq5      A.....A...  .A......A.  .T....TT.. .
seq6      AT.......   ........A.  ......T..T .
seq7      A.....A...  ........A.  ....T.T... .
seq8      A.....A.A.  ........A.  ......T... .
seq9      A.....A...  ....T...A.  .....TT... .
seq10     A.....A...  ........A. T.....T... T
seq11     ...T..A...  ..T...T.A.  ..T...T... .
seq12     A.....A...  ........A.  ......T... .
seq13     A.....A...  ........A.  ......T... .
seq14     A.....A...  ........A.  ......T..T T
seq15     A.....A.A.  .A......AT  .......T.. .
seq16     A.....A.AT  TA......A.  ......T... .
seq17     A.....A.A.  .A...TT.A.  ...TT...T. .
seq18     A.....A.A.  .A...T..A.  .......... .
seq19     A....TA.A.  .A......A.  .......... .
```

In addition to the 31 polymorphic sites, there are 19 fixed sites.
Mismatch dist: 1 4 11 27 33 33 28 26 12 6 6 3

Table 5.2: Polymorphic Sites from Data Set B

```
%          0123456789 0123456789
seq0       ATAAAATAAA TAAAATAAAA
seq1       .......... A.........
seq2       .......... A.........
seq3       .......... A.........
seq4       .......... AT..T.....
seq5       T.T.TT..TT A.T.....T.
seq6       T.T.TT..TT A.T.....T.
seq7       .A....AT.T A....A....
seq8       .A....AT.T A....A....
seq9       .A....AT.T A....A....
seq10      .A....AT.T A....A....
seq11      .A....AT.T A....A....
seq12      .A....AT.T A....A....
seq13      .A....AT.T A..T.A.T.T
seq14      .A.T.....T A....A....
seq15      .A.T.....T A....A....
seq16      .A.T.....T A....A....
seq17      .A.T.....T A....AT...
seq18      .A.T.....T A....AT...
seq19      .A.T.....T A....AT...
```

In addition to the 20 polymorphic sites, there are 30 fixed sites.
Mismatch dist: 25 12 3 25 27 30 15 12 9 3 9 18 0 0 2

# Homework 6

# Neutral Theory

Each exercise is worth 10 points.

To estimate separation times from genetic data, we rely on the "molecular clock." If, at some locus, two species differ by 0.01 (presumably neutral) nucleotide substitutions per site, and if neutral substitutions occur at a rate of $\rho$ per generation, then we estimate the separation time as $0.01/2\rho$.

**Exercise 6.1.** *Why is there a "2" in the denominator of this expression?*

To make such estimates, we must first estimate the rate, $\rho$, of neutral evolution. Such rates are estimated from the number of nucleotide substitutions that accumulate between DNA sequences separated for a known amount of time. Unfortunately, neither the number of substitutions nor the separation time is often known exactly, and this can add both uncertainty and bias to the molecular clock.

In the calculation just described, the number of nucleotide substitutions cannot be observed directly. What we do observe is the number of nucleotide differences. The problem is that when a substitution occurs at a nucleotide site that has previously mutated, it does not increase the number of site differences. The uncorrected number of site differences therefore underestimates the number of substitutions, a phenomenon known as *saturation*. This effect is insignificant in very recent comparisons but increases with age. It is exacerbated when nucleotide sites vary in rate, because rates at rapidly-evolving sites may be underestimated. In modern phylogenetic studies, these problems are addressed by fitting models of the substitution process [7, sec. 3.2–3.4]. If the model is appropriate, saturation adds noise but not bias to the molecular clock. Fortunately, estimates of dates are relatively insensitive to this component of the model: we get approximately the same answer from many different models [11, p. 143].

## 6.1 The Jukes-Cantor model of nucleotide substitution

Neutral theory predicts that substitutional changes will accumulate at a constant rate. Yet because of saturation, we cannot measure substitutional changes directly. What we do measure is the fraction, $p$, of sites that differ between each pair of sequences. We need a way to convert this into an estimate of the mean number, $K$, of substitutional changes per site. This requires some model of the substitutional process. A variety of such models have been introduced, but we will use only the simplest, which was introduced by Jukes and Cantor [5] in 1969.

Their model assumes that when a site mutates, it is equally likely to end up in any of the other three states. Under this assumption, Jukes and Cantor [5] show that

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \tag{6.1}$$

This formula allows us to estimate $K$ (the number of subsitutional changes per site) from $p$ (the fraction of sites that differ). If the neutral theory is correct, then $E[K] = 2ut$, where $u$ is the neutral mutation rate, and $t$ is the time since the last common ancestor of the two DNA sequences.

For example, consider the difference between humans and chimpanzees. We differ at about $35 \times 10^6$ of the $3 \times 10^9$ nucleotide sites in our (haploid) genomes—a fraction $p = 0.01167$. Plugging this into Eqn. 6.1 gives $K = 0.01177$ substitutions per site. There is hardly any difference between our estimates of $K$ and $p$, indicating that not much saturation has occurred. In cases such as this, there is no compelling reason to correct for saturation at all.

When $p$ is small, geneticists often use it as an approximation for $K$. This makes sense, provided that

the error involved in the approximation is small. How small is it in the chimpanzee-human comparison? Consider the relative error, which is defined as

$$\text{relerr} = \frac{|p - K|}{K}$$

where || indicates the absolute value. For the chimpanzee–human data, this gives relerr $= |0.01167 - 0.01177|/0.01177 \approx 0.008$. This means that when we use $p$ instead of $K$, we make an error that is less than 1% as large as the correct answer, $K$.

**Exercise 6.2.** *Calculate the relative error for a variety of values of p in order to decide when it is and is not necessary to correct for saturation.*

**Exercise 6.3.** *Fossils suggest that humans and chimpanzees last shared ancestor around* $t = 6 \times 10^6$ *years ago. The neutral theory implies that* $K = 2ut$*, where u is the neutral substitution rate. Use the chimpanzee-human data to estimate u. How does the estimate compare to the typical mammalian rate, approximately* $10^{-8}$ *per site per year? What might account for the difference?*

### 6.1.1 Deriving the Jukes-Cantor formula

Although we are interested in the process of mutation, we will focus here on an imaginary process of "perturbation." In each time unit, a site is perturbed with probability $\lambda$. After a perturbation the site is equally likely to be in any of the four states (A, T, G, and C). If it is perturbed, it ends up in a different state with probability 3/4. Consequently, mutations occur with probability $u = (3/4)\lambda$.

This formulation of the problem provides an enormous simplification: once a site has been perturbed, there is no further change in the probabilities of states A, T, G, and C. These probabilities are the same, no matter whether the site has been perturbed once, twice, or a thousand times. This makes the process far easier to study. Yet once we have derived a result, we can easily re-express it in terms of mutations rather than perturbations. All we'll need to do is replace $\lambda$ in our formula with the equivalent value $(4/3)u$.

Suppose that we compare homologous sites in two DNA sequences that have been separated for $t$ units of time. Because the rate of perturbation is constant, the number of perturbations along the the two branches that separate our DNA sequences is a Poisson random variable with mean $2\lambda t$. The probability that neither sequence has been perturbed is given by the zero term of the Poisson distribution: $e^{-2\lambda t}$. In this case, they are obviously identical. If there has been at least one perturbation, then the perturbed sequences are equally likely to be in any of the four states, and the two sites are identical with probability 1/4. The probability, $q$, of identity is therefore

$$\begin{aligned} q &= e^{-2\lambda t} + (1 - e^{-2\lambda t})/4 \\ &= \frac{1}{4} + \frac{3}{4} e^{-2\lambda t} \end{aligned}$$

Let us now re-express this result in terms of things we can measure. First substitute $(4/3)u = \lambda$, so that our formula is in terms of mutations rather than perturbations:

$$q = \frac{1}{4} + \frac{3}{4} e^{-(8/3)ut}$$

Next, re-express the formula in terms of the probability, $p = 1 - q$, that two homologous sites will differ:

$$p = \frac{3}{4} - \frac{3}{4} e^{-(8/3)ut}$$

Finally, substitute $K = 2ut$, the expected number of nucleotide site changes along the path that separates the two sites:

$$p = \frac{3}{4} - \frac{3}{4} e^{-(4/3)K}$$

After solving this equation for $K$, we end up with Eqn. 6.1.

## 6.2 Other models of substitution

Some nucleotide site changes (A↔G and C↔T) are especially common. These common changes are called "transitions," and the others are called "transversions." In mitochondrial DNA, transitions seem to be at least 30 times as common as transversions. With such data, the Jukes-Cantor model is clearly inappropriate, and various alternatives are often used.

One alternative is especially simple. If transitions are sufficiently common, relative to transversions, it may be reasonable to assume that *all* mutations are transitions. In such a model, some sites toggle back and forth between A and T, while others toggle between C and G. Either way, we can use

Table 6.1: Nucleotide sequence differences between complete mitochondrial genomes. Source: [2].

| | BN Rat | Mouse | Human |
|---|---|---|---|
| Brown Norway Rat | — | 4897 | 2897 |
| Mouse | | — | 5050 |
| Human | | | — |

a variant of the Jukes-Cantor model in which the number of states is two rather than four.

**Exercise 6.4.** *Derive a formula, analogous to Eqn. 6.1, in which the number of states is two rather than four. (Hint: repeat the steps in section 6.1.1. Your derivation should assume that, after a perturbation, the site is equally likely to be in either of the* two *states.)*

**Exercise 6.5.** *Use the 2-state model to estimate K from the chimpanzee-human data.*

## 6.3 Rodent mitochondrial DNA sequences

Table 6.1 shows the number of nucleotide site differences between the complete mitochondrial genomes of mouse, rat, and humans. These genomes are of slightly different sizes: about 16,300 bp for mouse and rat, and 16,569 bp for human.

**Exercise 6.6.** *Use the data in table 6.1 to estimate the fraction* $(p)$ *of site differences between each pair of species. To do this you will need to know the total number of sites in the mitochondrial genome. These numbers are about 16,300 bp for mouse and rat, and 16,569 bp for human. We can take the* smaller *of these two numbers as the effective size in our comparisons, because this is the largest number of sites we could conceivably align. Then (a) use the Jukes-Cantor formula to estimate K for each pair of species. (b) Because rats and mice are close relatives, we expect the rat-human number to equal the mouse-human number. How well do your results conform to this expectation?*

# Homework 7

# Selection

Many of the exercises in this homework have several components, which are given letters: *a*, *b*, and so on. Each component of each exercise is worth 10 points. Several of the exercises ask you to make graphs. Feel free to use any method you please. There is nothing wrong with a pencil and graph paper. And if you don't have graph paper, try this:

> For emergency graph paper, take out one sheet of ruled paper, turn it on its side, and place it beneath another sheet of ruled paper. If these two sheets have a light-colored backing—often provided by the rest of the pad or notebook—the vertical lines on the lower sheet are almost certain to show through well enough, combining with the horizontal lines on the top sheet to form a grid on which plotting is reasonably easy.
>
> Tukey [10, p. 43]:

## 7.1 How selection changes allele frequencies

**Exercise 7.1.** *At a biallelic locus, suppose that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have relative fitnesses 1, 1.02, and 1.03 and that the frequency of $A_1$ is $p = 0.1$. (a) What is the population's mean relative fitness? (b) What are the "marginal" or allele-specific relative fitnesses of $A_1$ and $A_2$? (c) What is the expected frequency of $A_1$ in the following generation?*

**Exercise 7.2.** *At a biallelic locus, suppose that $p = 0.2$, $s = 0.05$, and $h = 0.1$. Here, $p$ is the relative frequency of allele $A_1$, and the relative fitnesses of genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ are 1, $1 - hs$, and $1 - s$. Assume that the population is at Hardy-Weinberg equilibrium. (a) What is the population's mean relative fitness? (b) What are the "marginal" or allele-specific relative fitnesses of $A_1$ and $A_2$? (c) What is the expected frequency of $A_1$ in the following generation?*

**Exercise 7.3.** *Repeat exercise 7.2, assuming that $p = 0.5$, $s = 0.01$, and $h = -0.3$.*

**Exercise 7.4.** *Suppose that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have relative fitnesses 1, $1 - hs$, and $1 - s$, and that their absolute fitnesses are 1.3, 1, and 0.9. (a) What are $s$ and $h$? (b) Plot $\Delta_s p$ against $p$. (c) Where are the equilibria? Which are stable? Which are unstable? Hints: use Gillespie's Eqns. 3.2 and 3.3.*

**Exercise 7.5.** *Repeat exercise 7.4, assuming that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have absolute fitnesses 1.3, 1, and 1.2.*

**Exercise 7.6.** *Repeat exercise 7.4, assuming that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have absolute fitnesses 0.9, 1, and 0.8.*

**Exercise 7.7.** *Suppose that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have fitnesses 1, 1.15, and 1.2. The population is infinite and mates at random. What is the frequency of allele $A_1$ at the stable equilibrium? (No calculation is needed.)*

**Exercise 7.8.** *Suppose that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have fitnesses 1, 1.5, and 1. The population is infinite and mates at random. What is the frequency of allele $A_1$ at the stable equilibrium? (No calculation is needed.)*

**Exercise 7.9.** *Suppose that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have fitnesses 1, 0.5, and 1.5. The population is infinite and mates at random. What is the frequency of allele $A_1$ at the two stable equilibria? (No calculation is needed.)*

**Exercise 7.10.** *Use Gillespie's Eqn. 3.4 to graph $\hat{p}$ as a function of $h$ for $-1 < h < 2$. Your vertical axis should extend only from 0 to 1, because values outside this range are not legitimate*

*allele frequencies. Locate the regions of the graph that correspond to (a) overdominance, (b) incomplete dominance, and (c) underdominance. In making these determinations, assume that $s > 0$. For each of these regions, discuss the outcome of natural selection.*

**Exercise 7.11.** *Gillespie's equation 3.9 doesn't work if $h = 0$, i.e. if allele $A_2$ is completely recessive. Derive a formula for this case. Hints: (i) equation 3.6 still works; (ii) equation 3.8 simplifies to $\Delta_s p = pq^2 s/(1 - q^2 s) \approx pq^2 s$; (iii) it is still true that $\Delta p = \Delta_s p + \Delta_u p$.*

**Exercise 7.12.** *White flowers (genotype $rr$) are recessive to red ($RR$ and $Rr$) in an outbreeding plant species. In a large random sample, you count 200 white-flowered plants and 800 red-flowered plants. One generation later, you count 250 white and 750 red plants. To study these data, use the following notation for fitnesses and genotypic frequencies:*

| *G'type* | $RR$ | $Rr$ | $rr$ |
|---|---|---|---|
| *Fitness* | $w_{RR} = 1$ | $w_{Rr} = 1$ | $w_{rr} = 1 - s$ |
| *Freq* | $P_{RR} = p^2$ | $P_{Rr} = 2pq$ | $P_{rr} = q^2$ |

*This says that $p$ is the frequency of allele $R$, $q = 1-p$ is that of allele $r$, and the genotype frequencies are at Hardy-Weinberg equilibrium. Furthermore, relative fitness is 1 for both of the genotypes ($RR$ and $Rr$) that produce red flowers. It equals $1-s$ for the genotype ($rr$) that produces white flowers. To study the change between generations, we'll use the displayed equation at the top of Gillespie's p. 62. In the current notation, this equation becomes*

$$p' = \frac{p^2 w_{RR} + pq w_{Rr}}{\bar{w}} \tag{7.1}$$

*where $\bar{w} = p^2 w_{RR} + 2pq w_{Rr} + q^2 w_{rr}$ is the mean allele frequency. Answer the following questions: (a) What's the frequency, $p$, of allele $R$ in the first generation? (b) What's the frequency, $p'$, of allele $R$ in the second generation? (c) If this change was caused by selection, then what was the coefficient ($s$) of selection? This calculation is sensitive to numerical error: use at least 3 digits of precision throughout.*

**Exercise 7.13.** *Repeat exercise 7.12, this time assuming that the fraction of white flowers was $100/1000$ in the first generation and $95/1000$ in the second.*

*This calculation is sensitive to numerical error: use at least 3 digits of precision throughout.*

## 7.2 Selection and drift

In these problems, we'll follow the notation in Gillespie's section 3.9, which assumes that $A_1$ is the "wild type" allele, $A_2$ is the mutant allele, and genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have relative fitnesses 1, $1 + s/2$, and $1 + s$. (This notation differs from Kimura's, which Jon presented in lecture.)

**Exercise 7.14.** *Do we expect more adaptive evolution (fixations of advantageous mutations) in large or in small populations? Why?*

**Exercise 7.15.** *For a deleterious mutation with $s = -0.001$, what is the probability of fixation in a population of size (a) $N_e = 10,000$, (b) $N_e = 1000$, and (c) $N_e = 100$. Hint: In Python, Gillespie's equation 3.22 is*

```
# Prob of fixation of a new mutation.
def pfix(N, s):
    return (1-exp(-s))/(1-exp(-2*N*s))
```

**Exercise 7.16.** *Suppose that for the deleterious alleles of the preceding problem, the mutation rate is $u = 2.2 \times 10^{-9}$ per year. What is the rate of substitution of such alleles per million years?*

**Exercise 7.17.** *Problem 7.15 was about deleterious mutations. Repeat it for advantageous mutations with $s = 0.001$. (Hint: Use the* `pfix` *Python function defined above.)*

**Exercise 7.18.** *Repeat problem 7.16 for advantageous mutations with $s = 0.001$, still assuming that $u = 2.2 \times 10^{-9}$.*

On p. 93, Gillespie shows that if $s \ll 1 \ll 2N_e s$, and $N_e \approx N$, then the probability of fixation of a newly-arisen advantageous mutant allele is approximately

$$\pi_1(1/2N) \approx s \tag{7.2}$$

In other words, the probability of fixation for a new adaptive mutation is approximately twice the selective advantage of a heterozygote.

**Exercise 7.19.** *Repeat exercise 7.17, this time using the approximation in Eqn. 7.2. For each population size, calculate the relative error of the approximation,*

$$\text{relerr} = \left| \frac{s - \pi_1(1/2N)}{\pi_1(1/2N)} \right|$$

*where $\pi_1(1/2N)$ is the value given by Kimura's formula, and the vertical bars indicate the absolute*

*value.*[1] *Based on these relative errors, in which cases does the approximation work well, and in which does it work poorly?*

[1]For example, the absolute value of $-3$ is written $|-3|$ and equals 3.

# Homework 8

# Two Loci

**Exercise 8.1.** *Consider the following data set:*

| *Haplotype:* | *AB* | *Ab* | *aB* | *ab* |
|---|---|---|---|---|
| *Count:* | *30* | *70* | *50* | *20* |

*Use these data to answer the following questions: (a) What are the relative frequencies* $(x_1, x_2, x_3, x_4)$ *of the four gamete types* $(AB, Ab, aB, ab)$*? (b) What are the frequencies* $p_A$ *and* $p_B$ *of alleles A and B? (c) What is D? (d) What is the squared correlation,* $r^2$*, between loci? (Don't confuse this r with the recombination rate, which is also often called r. See Gillespie's p. 105.)*

**Exercise 8.2.** *Answer the same questions with these data:*

| *Haplotype:* | *AB* | *Ab* | *aB* | *ab* |
|---|---|---|---|---|
| *Count:* | *80* | *30* | *10* | *45* |

**Exercise 8.3.** *Answer the same questions as before, using the data below, and defining* $x_1$*,* $x_2$*,* $x_3$*, and* $x_4$ *to represent the relative frequencies of AG, AC, TG, and TC.*

| *Locus 1* | *Locus 2* |
|---|---|
| *A* | *G* |
| *A* | *G* |
| *A* | *C* |
| *T* | *G* |
| *T* | *G* |
| *T* | *G* |
| *T* | *G* |
| *T* | *G* |
| *T* | *C* |
| *T* | *C* |

**Exercise 8.4.** *Answer the same questions as before, using the data below, and defining* $x_1$*,* $x_2$*,* $x_3$*, and* $x_4$ *to represent the relative frequencies of AG, AC, TG, and TC.*

| *Locus 1* | *Locus 2* |
|---|---|
| *A* | *G* |
| *A* | *C* |
| *A* | *C* |
| *A* | *C* |
| *T* | *G* |
| *T* | *G* |
| *T* | *G* |
| *T* | *G* |
| *T* | *G* |
| *T* | *C* |

**Exercise 8.5.** *Suppose the recombination rate is* $c = 1/2$ *and that* $D = 0.2$ *in generation 0. What would D be after (a) 1 generation, and (b) 3 generations? (Hint: in last displayed equation on p. 104, Gillespie's r is my c.)*

**Exercise 8.6.** *Suppose the recombination rate is* $c = 1/1000$ *and that* $D = 0.3$ *in generation 0. What would D be after (a) 1 generation, and (b) 100 generations?*

**Exercise 8.7.** *D can be defined in several ways, including*

$$\begin{aligned} D &= x_1 - p_A p_B \\ &= x_1 x_4 - x_2 x_3 \end{aligned}$$

*Prove that these are equivalent. (Hint: Start by writing* $p_A$ *and* $p_B$ *in terms of the* $x_i$*s. Don't forget that* $\sum x_i = 1$*.)*

**Exercise 8.8.** *D can be defined in several ways, including*

$$\begin{aligned} D &= x_1 - p_A p_B \\ -D &= x_2 - p_A(1 - p_B) \end{aligned}$$

*Prove that these are equivalent. (Hint: remember that* $x_1$*,* $x_2$*,* $x_3$*, and* $x_4$ *are the frequencies of gametes AB, Ab, aB, and ab.)*

**Exercise 8.9.** *Suppose that, in generation 1, half the gametes are $A_1B_2$ and the other half are $A_2B_1$. What are the values of (a) the four gamete frequencies $(x_1, x_2, x_3, x_4)$ and (b) the coefficient, $D$, of linkage disequilibrium?*

**Exercise 8.10.** *Now assume that selection operates at the gamete stage, that alleles $B_1$ and $B_2$ are neutral, that allele $A_1$ has fitness $1 + s$ relative to $A_2$. What are (a) the four gametic fitnesses, $(w_1, w_2, w_3, w_4)$, and (b) the mean gametic fitness, $\bar{w}$. In calculating the mean, use the gametic relative frequencies $(x_1, x_2, x_3, x_4)$ that you calculated in problem 8.9. Your answers will be functions of a single unknown value, $s$.*

We have assumed that selection operates in the gamete stage. As explained in lecture, this model implies that the expected gamete frequencies in the following generation are

$$\begin{aligned} x_1' &= w_1(x_1 - cD)/\bar{w} \\ x_2' &= w_2(x_2 + cD)/\bar{w} \\ x_3' &= w_3(x_3 + cD)/\bar{w} \\ x_4' &= w_4(x_4 - cD)/\bar{w} \end{aligned}$$

where $\bar{w} = \sum_i x_i w_i$ is the mean fitness.

**Exercise 8.11.** *Assume that the recombination rate is $c = 1/4$, and take the other values $(w_i, \bar{w}, x_i, D)$ from your answers to problems 8.9 and 8.10. What are the expected values of the four gamete frequencies in the following generation? Your answers will be functions of a single unknown value, $s$.*

**Exercise 8.12.** *Soon after a selective sweep has finished, (a) what would you expect of nucleotide diversity at nearby loci? More diversity, less diversity, or about the same amount as before the sweep? (b) Is this effect larger or smaller in regions of low recombination?*

If two sites are 1 centimorgan (cM) apart on a chromosome, then the rate of recombination between them is $c = 1/100$. (This is the definition.) How does this relate to physical distance along the chromosome? In general, 1 cM is about a million base pairs—a "megabase"—for humans. In the exercise below, you will use this generalization to calculate the probability $\lambda$ of recombination between two adjacent nucleotides.

As a first approximation, you might suppose that

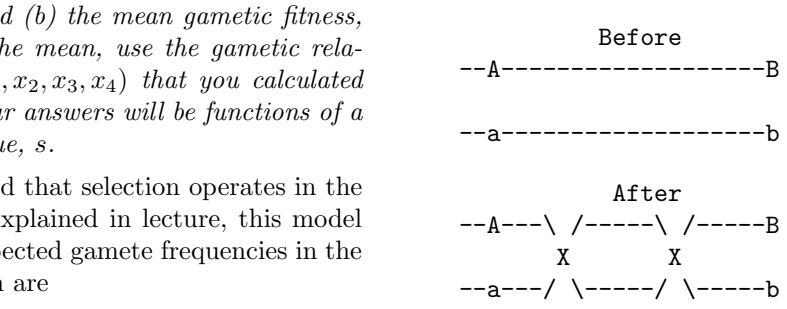
$$c = \lambda k$$

where $k$ is the physical distance between the two sites—the number of nucleotides between them. Our rule of thumb says that $c = 0.01$ when $k = 10^6$. Thus,

$$\lambda = 0.01/10^6 = 10^{-8}$$

The rate of recombination between adjacent nucleotides is about $10^{-8}$.

But this calculation is only approximate. To see why, consider sites A and B below.




In the "Before" picture, we see two chromosomes, one with the $AB$ haplotype, the other with $ab$. In the "After" picture, two cross-overs have happened, yet we still have the same two haplotypes. There were cross-overs but no recombination. This happens whenever the number of cross-overs between two loci is even. The probability ($c$) of recombination is the probability that the number of cross-overs is odd.

In 1919, JBS Haldane calculated the probability of an odd number of crossovers. His answer, now known as "Haldane's mapping function," is

$$c = (1 - e^{-2\lambda k})/2$$

provided that cross-overs occur at a constant rate along the chromosome.

**Exercise 8.13.** *Our rule of thumb (a centimorgan is a megabase) says that $c = 1/100$ when $k = 10^6$. Substitute these values into Haldane's formula above, and solve for $\lambda$. How does your result compare to the approximate one ($\lambda = 10^{-8}$) that we got above?*

**Exercise 8.14.** *Let's try extrapolating in the other direction. Suppose that two sites are separated by 50 megabases ($5 \times 10^7$ bases). Estimate $c$, assuming that $\lambda = 10^{-8}$, and using (a) the approximate formula, and (b) Haldane's mapping function. Finally, (c) comment of the accuracy of the approximation in this case.*

**Exercise 8.15.** *Make a graph. The horizontal axis represents $\lambda k$ and should run from 0 to 0.5.*

*On this graph, plot (a) the approximate formula $c \approx \lambda k$, and (b) Haldane's mapping function, as defined above. The vertical difference between the two lines shows the error involved in the approximate formula. (c) Over what range of $\lambda k$ is the approximation satisfactory? (d) Translate this range into units of base pairs (i.e. $k$), assuming that $\lambda = 10^{-8}$.*

On p. 110, Gillespie shows that a selective sweep removes variation at all sites such that (roughly) that $c/s < 0.1$, where $c$ is the recombination rate with the selected site. (This applies both upstream and downstream from the selected site.)

**Exercise 8.16.** *How large is the affected region of the chromosome if $\lambda = 10^{-8}$ and $s = 0.001$?*

**Exercise 8.17.** *How large is the affected region of the chromosome if $\lambda = 10^{-8}$ and $s = 0.1$?*

**Exercise 8.18.** *In Europeans, the allele for lactase persistence sits on a region of LD that extends for a megabase. The persistence allele is still polymorphic, but let us suppose that it sweeps to fixation, removing most of the variation from this entire megabase-sized region. Based on your answer to the previous question, how strong would this selection need to be?*

**Exercise 8.19.** *Identify and describe one recombination event in table 8.1. (Don't use any of the events described in the answered questions in the back.)*

**Exercise 8.20.** *Identify and describe another recombination event in table 8.1. (Don't use any of the events described in the answered questions in the back.)*

Table 8.1: DNA sequences near the human lactase gene, typed in a European sample. Columns are nucleotide sites. Top row (the reference sequence) shows common state at each site. Capital "A" is the lactase persistence allele. Dots indicate identity with top row.

```
   1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1
   aaggaggcga cattccgctt caggcattcc tatctaaaca gaccaacgta Agggtacaat gcctaaccca gacgtttcaa ctctggctgt tattcctcga t
01 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
02 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
03 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
04 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
05 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
06 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
07 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
08 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
09 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
10 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
11 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
12 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
13 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
14 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
15 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
16 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
17 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
18 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
19 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
20 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
21 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
22 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
23 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
24 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
25 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
26 .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .
27 .......... .......... .......... ...t...... .......... .......... .......... .......... .......... .......... .
28 .......... .......... .......... ...t...... .......... .......... .......... .......... .......... .......... .
29 .......... .......... .......... .......... .......... .......... .....c.... .......... .......... .......... .
30 .......... .......... .......... .......... .......... .......... ....g..... .......... .......... .......... .
31 gg.a.ca.ag .g.gt..... .......... .......... .......... .......... .......... .......... .......... .......... .
32 .......... .......... .......... .......... .......... G......... .......... .......... .......... .......... .
33 .......... .......... .......... .......... .......... G......... .......... .......... .......... .......... .
34 .......... .......... .......... .......... .......... G......... .......... .......... .......... .......... .
35 .......... .......... .......... .......... .......... G......... .......... .......... .......... .......... .
36 .......... .......... .......... .......... .......... G......... .......... .......... .......... .......... .
37 .......... .......... .......... .......... .......... G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
38 .......... ..c.....cc gga....gat ..at..gg.. c.....tc.g Gaaa.g..cc ttt...tg.. ....c...t. t......... .g.....t.. .
39 ....g....g ..c.....cc gga....gat ..at..gg.. c.....tc.g Gaaa.g..cc ttt...tg.. ....c...t. t......... .g.....t.. .
40 gg.a..at.g t.c.t..tcc ...agtag.t .cat..g... ..t..ttccg G..a.gt... ..t....... ..gac.c.tg tct....... .......... .
41 .......... ..c.t..tcc ...agtag.t .cat..g... ..t.gttccg G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
42 .......... ..c.t..tcc ...agtag.t .cat..g... ..t.gttccg G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
43 ..aa..at.g t.c.t..tcc ...agtag.t .cat..g... ..t.g.tc.g G..a.gt... ..t....... ..gac.c.tg tct....... .g...t.t.. c
44 ..aa..at.g t.c.t..tcc ...agtag.t .cat..g... ..t..ttc.g G..acgt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
45 ..aa..at.g t.c.t..tcc ...agtag.t .cat..g... ..t.gttc.g G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
46 ....g....g ..c.....cc gga....gat ..at..gg.. c.....tc.g Gaaa.g..cc ttt...tg.. ....cg.gt. t..ctata.c cg.c..ctcg .
47 gg.a..at.g t.c.t..tcc ...agtag.t .cat..g... ..t.gttccg G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
48 gg.a..at.g t.c.t..tcc ...agtag.t .cat..g... ..t.gttccg G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
49 gg.a..at.g t.c.t..tcc ...agtag.t .cat..g... ..t.gttccg G..a.gt... ..t....... ..gac.c.tg tct.....a. .gc..t.t.. c
50 .........g ..c..tatcc gga....g.t c.atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
51 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ct...a. .gc..t.t.. c
52 gg.a.ca.ag .g.gtta.cc gga....g.t ..atc.g.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
53 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
54 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
55 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
56 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
57 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg tg..cg.gt. t..ctata.c cg.c..ctcg .
58 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.ca.ctcg .
59 gg.a.ca.ag .g.gtta.cc gga....g.t c.atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg t...cg.gt. t..ctata.c cg.c..ctcg .
60 gg.a.ca.ag .g.gtta.cc gga....g.t ..atcgg.tc .g.tg.tc.g G..a.g.g.. ..tg....gg tg..cg.gt. t..ctata.c cg.c..ctcg .
   1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1
```

# Homework 9

# Inbreeding

Each exercise is worth 10 points.

**Exercise 9.1.** *Find $p$ and $F$ for a sample with genotype frequencies 100, 200, and 200 of genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$, respectively.*

**Exercise 9.2.** *Find $p$ and $F$ for a sample with genotype frequencies 200, 100, and 200 of genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$, respectively.*

**Exercise 9.3.** *At a bi-allelic locus, alleles $A_1$ and $A_2$ have frequencies $p = 1/4$ and and $q = 3/4$. Relative to the current generation, $F$ is $1/2$. What is the frequency of genotype $A_1A_1$?*

**Exercise 9.4.** *At a bi-allelic locus, alleles $A_1$ and $A_2$ have frequencies $p = q = 1/2$. Relative to the current generation, $F$ is $1/4$. What is the frequency of genotype $A_1A_1$?*

**Exercise 9.5.** *What is the coefficient of kinship of half-siblings?*

**Exercise 9.6.** *What is the coefficient of kinship of full cousins?*

**Exercise 9.7.** *Figure 9.1 shows the genealogy of George Darwin, one of Charles Darwin's sons. What is his inbreeding coefficient?*

**Exercise 9.8.** *Suppose that (a) we are studying a locus with two alleles, $A_1$ and $A_2$, (b) the frequency of allele $A_1$ is $p = 0.4$, and (c) the coefficient of kinship between two individuals is 1/8. What are the probabilities that their offspring will be (1) an $A_1A_1$ homozygote, (2) an $A_1A_2$ heterozygote, and (3) an $A_2A_2$ homozygote?*

**Exercise 9.9.** *Suppose that (a) we are studying a locus with two alleles, $A_1$ and $A_2$, (b) the frequency of allele $A_1$ is $p = 0.1$, and (c) the coefficient of kinship between two individuals is 1/16. What are the probabilities that their offspring will be (1) an $A_1A_1$ homozygote, (2) an $A_1A_2$ heterozygote, and (3) an $A_2A_2$ homozygote?*

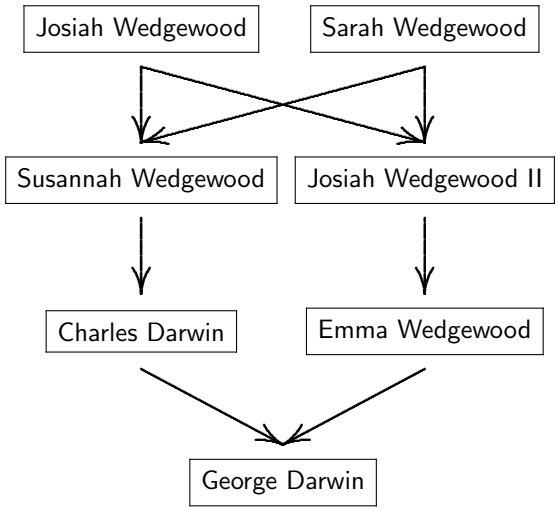



Figure 9.1: Genealogy of George Darwin

**Exercise 9.10.** *This exercise is the same as the last one, except that this time you are not given numerical values for the various parameters. This time, you know only that (a) the frequency of allele $A_1$ is $p$, and (b) the coefficient of kinship between two individuals is $f$. What are the probabilities that their offspring will be (1) an $A_1A_1$ homozygote? (2) an $A_1A_2$ heterozygote? (3) an $A_2A_2$ homozygote? Your answers should be formulas, not numbers.*

# Homework 10

# Population Structure

$F_{ST}$ can be expressed in terms of gene diversity:

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{10.1}$$

in terms of homozygosity (Gillespie's formula 5.3, p. 132):

$$F_{ST} = \frac{G_S - G_T}{1 - G_T} \tag{10.2}$$

or in terms of the variance $V$ between allele frequencies of subdivisions:

$$F_{ST} = V/pq \tag{10.3}$$

**Exercise 10.1.** *Verify algebraically that formulas 10.1 and 10.2 are equivalent. (Hint: Use the facts that $G_S = 1 - H_S$ and $G_T = 1 - H_T$. This just says that every genotype is either a homozygote or a heterozygote, so homozygosity and heterozygosity must sum to 1.)*

**Exercise 10.2.** *Verify that formulas 10.1 and 10.3 are equivalent. (Hint: Use Wahlund's formula, which was explained in the lecture on population structure. Or consult Gillespie's discussion of $F_{ST}$.)*

In 1978, Sewall Wright published a chapter on variation among human populations. He included a table of allele frequencies at six loci, from which I have extracted data on a two biallelic loci. (I also left out several of his populations and changed his racial labels to geographic ones.) The table below shows the frequencies of the $P^1$ allele at the $P$ locus and the $Fy$ allele at the Duffy locus:

| Population | $P^1$ | $Fy$ |
|---|---|---|
| African | 0.734 | 0.037 |
| European | 0.496 | 0.436 |
| Australian | 0.330 | 1.000 |
| Asian | 0.259 | 0.901 |
| Average | 0.455 | 0.593 |

**Exercise 10.3.** *What is $F_{ST}$ at the $P$ locus?*

**Exercise 10.4.** *What is $F_{ST}$ at the Duffy locus?*

**Exercise 10.5.** *What is the average of the two $F_{ST}$ estimates? How does your estimate compare with the value of 0.12, which has been found in many other studies of genetic differences between continental human populations?*

**Exercise 10.6.** *Suppose that (1) a population is divided into several groups, within which mating is random, (2) the frequency of allele $A$ is $\bar{p} = 1/4$ in the population as a whole, and (3) $F_{ST} = 1/3$. What is the expected frequency within the population as a whole of genotypes $AA$, $Aa$, and $aa$?*

**Exercise 10.7.** *Suppose that (1) a population is divided into several groups, within which mating is random, (2) the frequency of allele $A$ is $\bar{p} = 1/2$ in the population as a whole, and (3) $F_{ST} = 1/10$. What is the expected frequency within the population as a whole of genotypes $AA$, $Aa$, and $aa$?*

Sewall Wright showed that at equilibrium between migration and genetic drift,

$$E[F_{ST}] = \frac{1}{4Nm + 1}$$

where $N$ is the diploid effective size of each subpopulation. The result is based on Wright's "Island Model" of population structure, which assumes there is an infinite number of sub-populations, and that in each generation a fraction $m$ of each subpopulation consists of immigrants sampled from the population as a whole. This model is unrealistic for several reasons, among them being the infinite number of sub-populations and the equal rate of exchange between each pair of sub-populations. In the question below, ignore these discrepancies.

**Exercise 10.8.** *Among continental human populations, estimates of $F_{ST}$ are usually close to*

1/9. *Assume that this represents an equilibrium between migration and genetic drift under the Island Model of population structure. What does this imply about the number $Nm$ of migrants between pairs of populations in each generation?*

# Homework 11

# Admixture

**Exercise 11.1.** *The table below summarizes a comparison among four haploid genomes: one European, one African, one Neanderthal, and one chimpanzee. "0" represents the ancestral allele and "1" the derived (i.e. mutant) allele. Each column refers to a different pattern in which these alleles may occur at an individual nucleotide site. The bottom row shows the number of nucleotide sites that have each pattern.*

| | Nucleotide Site Pattern | | |
|---|---|---|---|
| | *xy* | *yn* | *xn* |
| African (X) | 1 | 0 | 1 |
| European (Y) | 1 | 1 | 0 |
| Neanderthal (N) | 0 | 1 | 1 |
| Chimpanzee (C) | 0 | 0 | 0 |
| | 303,340 | 103,612 | 95,347 |

*Explain why, in the absence of gene flow from the Neanderthal population into modern humans, we expect site patterns yn and xn to be approximately equal in frequency.*

# Homework 12

# Quantitative Characters

Table 12.1: Quinoa seed weights. Each column refers to a different seed weight, each row to the sample of seeds from a different inbred line of quinoa. The entries give the number of seeds of each weight observed in each line

| weight: | 2 mg | 3 mg | 4 mg | 5 mg | 6 mg |
|---|---|---|---|---|---|
| line 1 | 3 | 32 | 38 | 22 | 5 |
| line 2 | 1 | 19 | 24 | 43 | 13 |
| line 3 | 6 | 23 | 41 | 26 | 4 |
| line 4 | 12 | 38 | 35 | 13 | 2 |
| Sum | 22 | 112 | 138 | 104 | 24 |

You are studying the quantitative genetics of seed weight in quinoa (*Chenopodium quinoa*), an extremely nutritious non-cereal grain of the Andes. The seeds have very high protein content, with a good balance of the essential amino acids. According to Wikipedia, "quinoa is being considered a possible crop in NASA's Controlled Ecological Life-support System for long-duration manned spaceflights." You grow four inbred lines under uniform conditions in the same field and weigh 100 randomly chosen seeds from each line. The table summarizes your data. You want to know how much heritable variation for seed weight there is.

Several of the exercises below will ask you to estimate the mean and variance. These are estimated as

$$\bar{X} = S_1/N \tag{12.1}$$

$$V = \frac{S_2 - S_1^2/N}{N-1} \tag{12.2}$$

where $N$ is the number of observations, $S_1$ is the sum of the observed values, and $S_2$ is the sum of squares of observed values. With ordinary (ungrouped) data,

$$S_1 = \sum_{i=1}^{N} x_i, \quad \text{and}$$

$$S_2 = \sum_{i=1}^{N} x_i^2$$

Here, $x_i$ is the value of the $i$th observation, and the sums run across all $N$ items in the data set. Our data, however, are grouped, so we can take a shortcut. Let $x$ represent one of the values that the data may take. In table 12.1, for example, $x$ takes the values 2, 3, 4, 5, and 6. The number of observations with value $x$ is written as $n_x$. For example, in the data for line 1, $n_2 = 3$, $n_3 = 32$, and $n_4 = 38$. With such data,

$$N = \sum_x n_x,$$

$$S_1 = \sum_x x n_x, \quad \text{and}$$

$$S_2 = \sum_x x^2 n_x$$

Now the sums run across the 5 values that data items may take rather than the 100 items in each data set. Use whichever method you prefer in calculating $N$, $S_1$, and $S_2$. Then use equations 12.1 and 12.2 to estimate the mean and variance.

**Exercise 12.1.** *Calculate the mean and variance of seed weight for line 1.*

**Exercise 12.2.** *Calculate the mean and variance of seed weight for line 2.*

**Exercise 12.3.** *Calculate the mean and variance of seed weight for line 3.*

**Exercise 12.4.** *Calculate the mean and variance of seed weight for line 4.*

**Exercise 12.5.** *Calculate the mean and variance of seed weight for all lines taken together. By the way, this variance estimates $V_P$, the total phenotypic variance.*

**Exercise 12.6.** *Calculate the variance of the four line means. Because the sample from each line is pretty large, this variance is influenced only slightly by environmental effects. It is almost pure genetic variance. We will take it as an estimate of $V_G$.*

**Exercise 12.7.** *Calculate the mean of the four within-line variances. This is an estimate of the variance within lines. Because the lines are inbred, there is no genetic variation within them, and the variance you calculate here estimates the environmental component of variance, $V_E$.*

**Exercise 12.8.** *In exercises 12.6 and 12.7, you estimated the genetic component of variance, $V_G$, and the environmental component, $V_E$. The sum of these should equal the phenotypic variance, $V_P$, which you estimated in question 12.5. Does it?*

**Exercise 12.9.** *What fraction of the phenotypic variance ($V_P$) is accounted for by $V_G$? In other words, what is $V_G/V_P$?*

**Exercise 12.10.** *What kind of heritability estimate did you calculate in exercise 12.9? Narrow-sense or broad-sense?*

Encouraged that you have heritable variation for seed weight, you outcross the plants and apply a selection differential ($S$) of 1 mg to the seeds, after harvesting them in bulk from all of the plants in your field. After four generations, the mean seed weight has increased by 0.2 mg.

**Exercise 12.11.** *What is the mean seed weight now? What is the realized (narrow-sense) heritability? Why does it differ from your previous estimate (question 12.9)?*

**Exercise 12.12.** *Approximately, what is the additive genetic variance ($V_A$) of seed weight in your study population? What is the dominance variance, $V_D$? (Assume there is no other form non-additive genetic variance besides the dominance variance). Why are you disappointed?*

You decide to test your new estimate of the narrow-sense heritability by estimating the regression of offspring seed weights on parental seed weights. This is a lot of work, but you are obsessed. You record the weights of seeds chosen randomly from the population, then grow them individually in pots and cross-fertilize them. When their seeds are fully mature, you weigh samples of seeds from each plant. This allows you to plot the weights of offspring seeds against the average weights of their parents (when they were seeds).

Table 12.2: Data from Pearson and Lee [8]: the correlation between mothers and daughters and the phenotypic variance.

| | Stature | Span | Cubit |
|---|---|---|---|
| Mother-daughter corr. | 0.507 | 0.452 | 0.421 |
| Phenotypic var. ($in^2$) | 6.26 | 8.27 | 0.784 |

**Exercise 12.13.** *If you were to plot offspring values against midparent values and fit a linear regression to these data, what numerical value would you expect the slope to have?*

What if you had selected on a whole-plant basis, rather than on an individual-seed basis? You worked at the level of individual seeds because the seeds are highly variable (differing by more than a factor of three, from smallest to largest). But most of that variation is environmental, so of course the seed-to-seed heritability is low. If you averaged all the seeds on each plant, then the mean seed weights of (adult) plants would undoubtedly be more heritable (i.e., show a much larger value of $h^2$) than the weights of individual seeds. This suggests that you might make more progress (per generation) in your quest to increase seed weights, if you selected on the mean seed weights of whole plants, rather than on the weights of individual seeds. But maybe not!

**Exercise 12.14.** *Why might this approach be just as slow as the approach you took? Hint: The rate of evolution depends on the absolute amount of additive genetic variation for the trait, not just on $h^2$!*

In 1903, Pearson and Lee [8] published the results of what was then the most extensive study ever of the inheritance of human physical characteristics. These data occupy a distinguished position in the history of population genetics, since they were the basis of R.A. Fisher's 1918 famous demonstration that Mendelian inheritance could account for variation in continuous characters. Pearson and Lee collected data on stature, span (distance between fingertips of outstretched arms), and cubit (fore-arm length), from about 1100 families. Some of these data are shown in table 12.2.

**Exercise 12.15.** *Use the data of Pearson and*

*Lee to estimate the heritability of stature. (Hint: use Gillespie's table 6.2.)*

**Exercise 12.16.** *Use the data of Pearson and Lee to estimate the heritability of span. (Hint: use Gillespie's table 6.2.)*

**Exercise 12.17.** *Use the data of Pearson and Lee to estimate the heritability of cubit. (Hint: use Gillespie's table 6.2.)*

**Exercise 12.18.** *Over the past 100 years, the stature of young women in most developed countries has increased about 0.157 inches per decade. This is called the "secular trend" in stature. This trend probably reflects changes in diet and/or public health. But let us consider the possibility that it was caused by natural selection. What selection gradient ($\beta$) does this hypothesis imply? You'll need to convert the change per decade into change per generation. For this purpose, assume that a human generation is 28 years.*

**Exercise 12.19.** *The average bill depth of Geospiza fortis (Darwin's medium ground finch) increased by 0.5 mm in one generation (1976 to 1978) in the population on Isla Daphne Major that has been studied by Peter and Rosemary Grant and their students and colleagues since the early 1970s. The phenotypic standard deviation of this trait is around 1 mm, and its heritability has been estimated (from the correlations among relatives) to be around $h^2 = 0.8$.*

*Given these numbers, what was the selection gradient ($\beta$) during the drought of 1977? (The drought was caused by a severe El Niño event that forced the birds to feed on large, hard seeds that they otherwise wouldn't eat.)*

# Appendix A

# Answers

**Answer *1.2.*** 1/4

**Answer *1.4.*** 1/2

**Answer *1.6.*** There are 6 ways to choose 2 out of 4: $\binom{4}{2} = \frac{4!}{2!\times 2!} = 24/[2\times 2] = 6$. They are: HHTT, HTHT, HTTH, THHT, THTH, and TTHH.

**Answer *1.8.*** $6 \times 1/2^4 = 6/16 = 3/8$

**Answer *1.10.*** $1/6 + 1/6 = 1/3$.

**Answer *1.12.*** $1/2 \times 2/5 = 1/5$

**Answer *1.14.*** There are two ways to answer this question. The first is algebraic: point out that $X$ is a binomial distribution with parameters $N = 3$ and $p = 1/2$. This implies that $X$ can take values 0, 1, 2, and 3; it takes value $x$ with probability $P_x = \binom{3}{x}/8$. The second way to answer the question is to write the answer in tabular form:

| $x$ | $P_x$ |
|---|---|
| 0 | 1/8 |
| 1 | 3/8 |
| 2 | 3/8 |
| 3 | 1/8 |

**Answer *1.16.*** The hard way to answer this question is to evaluate $V[X] = E[X^2] - E[X]^2$, or some similar expression. The easy way is to point out that, because $X$ is binomial with $N = 3$ and $p = 1/2$, the the variance is $V[X] = Np(1-p) = 3/4$.

**Answer *1.18.*** Freq. of rainy: $100/200 = 1/2$

**Answer *1.20.*** Freq. of rainy given sad: $70/80 = 7/8$

**Answer *1.22.*** $E[Y] = 2p$ and $V[Y] = 2p(1-p)$. There are two simple ways to get these results:

1. The question implies that $Y$ is binomial with $N = 2$ and probability parameter $p$. Consequently, $E[Y] = 2p$ and $V[Y] = 2p(1-p)$.
2. The question says that $Y$ is a sum of two independent values, each of which is a Bernoulli random variable with mean $p$ and variance $p(1-p)$. Therefore, $E[Y] = 2p$ and $V[Y] = 2p(1-p)$.

There are also harder ways, which involve deriving the properties of the Binomial or Bernoulli distributions.

**Answer *1.24.*** Because $X$ is binomial with $N = 2$ and probability parameter $p = 0.8$, its mean is $E[X] = 2p = 1.6$.

**Answer *1.26.*** Answer not provided.

**Answer *1.28.*** Pr one death: $0.61e^{-0.61}$

**Answer *1.30.*** Pr 2 deaths: $0.61^2e^{-0.61}/2! = 0.101$

**Answer *1.32.*** Pr 1 child: $2e^{-2} = 0.27$.

**Answer *1.34.*** This is exactly like Fig. 2 of JEPr, which presents the following probability distribution:

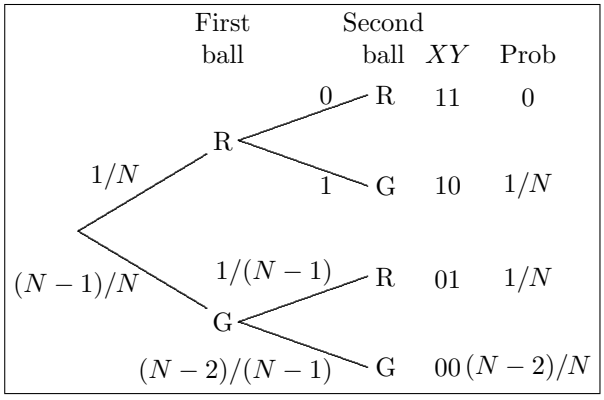
| Event | Prob |
|---|---|
| RR | 1/6 |
| RG | 1/3 |
| GR | 1/3 |
| GG | 1/6 |

where "R" and "G" stand for "red" and "green", "RG" means "1st ball red and 2nd green," and so on. For the current question, we translate "R" and "G" into into "1" and "0," and we define $X$ and $Y$ to equal the values of ball 1 and ball 2. The probability distribution becomes

| $X$ | $Y$ | $\Pr(X,Y)$ |
|---|---|---|
| 1 | 1 | 1/6 |
| 1 | 0 | 1/3 |
| 0 | 1 | 1/3 |
| 0 | 0 | 1/6 |

**Answer *1.36.*** $E[X^2] = E[Y^2] = 1/2$, so $V[X] = E[X^2] - E[X]^2 = 1/2 - 1/4 = 1/4$. $V[Y]$ is also =1/4.

**Answer *1.38.***




**Answer *1.40.*** The probability distribution of $X$ and $Y$ is shown on the right side of the figure in the answer to exercise 1.38. For reference, it looks like this:

| $X$ | $Y$ | Pr |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 0 | $1/N$ |
| 0 | 1 | $1/N$ |
| 0 | 0 | $(N-2)/N$ |

Using this distribution,

$$\begin{aligned} E[Y] &= 1 \times 0 \\ &\quad + 0 \times 1/N \\ &\quad + 1 \times 1/N \\ &\quad + 0 \times (N-2)/N \\ &= 1/N \end{aligned}$$

This is also $E[Y^2]$, because $Y$ takes only the values 0 and 1, and $Y^2 = Y$. Thus, $V[Y] = E[Y^2] - E[Y]^2 = 1/N - 1/N^2 = (1/N)(1 - 1/N)$. An alternative answer: these results follow because $Y$ is a Bernoulli random variable with parameter $p = 1/N$.

**Answer *1.42.*** 2/5 or 0.4

**Answer *1.44.*** normal

**Answer *1.46.*** binomial

**Answer *2.2.*** $P_{CD} = 15/100 = 0.15$

**Answer *2.4.*** By gene counting, $p_C = 175/200 = 0.875$. By the formula, $p_C = 0.8 + 0.15/2 = 0.875$.

**Answer *2.6.*** There is more than one correct answer. One approach begins with the observation that the frequency of $A_1$ is the same as the probability that a random gene copy chosen from a random individual is allele $A_1$. Let us calculate this probability.

With probability $P_{11}$, we choose genotype $A_1A_1$. In this case, we get allele $A_1$ with probability 1. With probability $P_{12}$, our individual is $A_1A_2$, and we then get $A_1$ with probability 1/2. If we choose $A_2A_2$, we cannot possibly get $A_1$. Thus, the probability of $A_1$ is

$$(P_{11} \times 1) + (P_{12} \times 1/2) + (P_{22} \times 0) = P_{11} + P_{12}/2$$

**Answer *2.8.*** For allele C, $p = 0.15 + 0.5/2 = 0.4$. For T, $q = 1 - p = 0.6$

**Answer *2.10.*** $H_{obs} = 0.5$.

**Answer *2.12.*** Using (2.2), $F = (P_{11} - p^2)/pq = (1/3 - 1/4)/(1/4) = 1/3$. Using (2.3), $F = -(P_{12} - 2pq)/2pq = -(1/3 - 1/2)/(1/2) = 1/3$. Using (2.4), $F = (P_{22} - q^2)/pq = (1/3 - 1/4)/(1/4) = 1/3$.

**Answer *2.14.*** $P_{11} = 0.726027$, $P_{12} = 0.260274$, $P_{22} = 0.013699$, $p_1 = 0.856164$, and $F = -0.056762$.

**Answer *3.2.***

$$\frac{1}{2} \times \left(1 - \frac{1}{10}\right) = \frac{9}{20} = 0.45$$

**Answer *3.4.*** $2N = 17.93$.

**Answer *3.6.*** Same as above. We start with formulas like this:

$$\begin{aligned} H_9 &= H_0(1 - 1/26)^9 \\ H_{18} &= H_9(1 - 1/10)^9 \end{aligned}$$

Substituting $H_9$ into the second equation gives

$$H_{18} = H_0(1 - 1/26)^9(1 - 1/10)^9$$

You can multiply in any order, so the answer to the second problem is the same as that of the first.

**Answer *3.8.***

$$t = \frac{\ln(H_t/H_0)}{\ln(1 - 1/2N)}$$

**Answer *3.10.*** $4Nu = 36 \times 10^{-6}$, and the expected heterozygosity is $4Nu/(1 + 4Nu)$, or 0.000036. The expected number of heterozygous flies is 16 times this number, or 0.00058, which rounds to 0. You would probably see no heterozygous flies.

**Answer *3.12.*** Heterozygosity is maximal when $\theta \to \infty$. For the model of infinite sites, this gives $\max \hat{H} = 1$.

**Answer *3.14.*** The calculation is based on Eqn. 3.3, because we are using the symmetric biallelic model of mutation. The question tells us that

$k = 2$ and $\hat{H} = 0.005$. Eqn 3.3 becomes $0.005 = \theta/(1 + 2\theta)$, or $\theta = 0.00505$. And since $\theta = 4Nu$, where $u = 2 \times 10^{-7}$, we have $2N = 12626$.

**Answer *4.2.*** For an interval with $i$ lines of descent, the expected duration in generations is $4N/[i(i-1)]$. For our problem, $2N = 5000$. The expected duration is therefore. (a) 5000 for 2 lines of descent, (b) $1000/9 \approx 111$ for 10, and (c) $10000/(1000 \times 999) = 10/999 \approx 0.01$ for 1000.

**Answer *4.4.*** We have 2 coalescent intervals, one with 3 lines of descent and the other with 2. The expected duration of the one with 3 lines of descent is 5000/3 and the expected duration of the other is 5000 generations. The expected depth of the tree is the sum of these, $5000 + 5000/3 = 20000/3$. The easy way to get this answer is with the formula $4N(1 - 1/K)$, where $K = 3$ is the number of gene copies in the modern sample. This gives the same answer, $10000 \times 2/3 = 20000/3$.

**Answer *4.6.*** For an interval with $i$ lines of descent, the expected length in generations is $4N/[i(i-1)]$, and the total branch length within the interval is $i$ times this value, or $4N/(i-1)$. The expected number of mutations is $u$ times the total branch length, or $\theta/(i-1)$, where $\theta = 4Nu$. For our problem, $\theta = 10$. The expected numbers of mutations are therefore (a) $10/1 = 10$ for 2 lines of descent, (b) $10/9 = 1.11$ for 10, and (c) $10/999 = 0.01$ for 1000.

**Answer *4.8.*** 18.3

**Answer *4.10.*** 0.797

**Answer *4.12.*** Mean pairwise diff: $\pi = 11.69$ per sequence, and $\pi = 0.29$ per site.

**Answer *4.14.*** $S = 30$ per sequence or 0.75 per site

**Answer *4.16.*** $\hat{\theta}_\pi = 11.69$ per sequence or 0.29 per site.

**Answer *4.18.*** $\hat{\theta}_S = 30/2.83 = 10.6$ per sequence or $0.75/2.83 = 0.265$ per site.

**Answer *4.20.*** In per-site units, we need to compare $\hat{\theta}_\pi = 0.29$ and $\hat{\theta}_S = 0.265$. These numbers don't differ too much, so there is no obvious reason to reject the model of neutral DNA in a randomly mating population of constant size.

**Answer *4.22.*** Table A.1 presents data set B again, with an extra row at the top showing the minor allele counts: Tabulating the counts gives

| Minor allele count | Number of sites |
|---|---|
| 1 | 9 |
| 2 | 9 |
| 3 | 6 |
| 4 | 5 |
| 5 | 1 |

**Answer *4.24.*** The expected numbers are $\hat{\theta}(1 + 1/9) = 11.78$ for singletons, $\hat{\theta}(1/2 + 1/8) = 6.63$ for doubletons, $\hat{\theta}(1/3 + 1/7) = 5.05$ for tripletons, $\hat{\theta}(1/4 + 1/6) = 4.41$ for quadrupletons, and $\hat{\theta} \times 1/5 = 2.12$ for quintupletons.

**Answer *4.26.*** Poisson with mean 3.

**Answer *4.28.*** $\theta = 0.01$ for Europe and Asia; 0.0204 for Africa.

**Answer *4.30.*** Each segregating site (whether a singleton or not) increments the value of $S$ by 1. Consequently, it increments $\hat{\theta}_S$ by $\left(\sum_{k=1}^{K-1} 1/k\right)^{-1}$.

**Answer *4.32.*** Many segregating sites will be singletons, which add more to $\hat{\theta}_S$ than to $\hat{\theta}_\pi$.

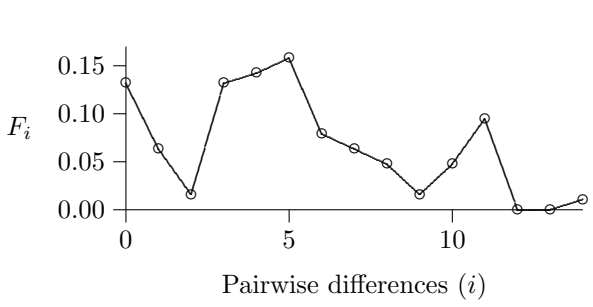
**Answer *5.2.*** The expanded population should have an excess of sites in which the derived allele is present in only one copy: an excess of singletons.

**Answer *5.4.*** The sum of the mismatch distribution is 45, which is also equal to $10 \times 9/2$.

**Answer *5.6.***

$$\pi = \Big(\overbrace{2\times 8}^{\text{site 0}} + \overbrace{4\times 6}^{\text{site 1}} + \overbrace{2\times 8}^{\text{site 2}} + \overbrace{2\times 8}^{\text{site 3}} + \overbrace{3\times 7}^{\text{site 4}} + \overbrace{2\times 8}^{\text{site 5}} + \overbrace{1\times 9}^{\text{site 6}} + \overbrace{1\times 9}^{\text{site 7}} + \overbrace{4\times 6}^{\text{site 8}} + \overbrace{3\times 7}^{\text{site 9}}\Big)/45 = 172/45 = 3.82$$

**Answer *5.8.***




The vertical axis ($F_i$) is the fraction of pairs of DNA sequences that differ by the number of sites shown on the horizontal axis.

Table A.1: Data set B with counts of minor allele at top

```
      35 43  22  23112123 2 42 4311142 14 213  1 <-counts
seq01 TGCCACTCCA ATCTCTCGCC AGATGGCATG CCTTATCGCG
seq02 .......... G......... .A.C...GCA T.........
seq03 .......... G......A.. .A.C...GC. T....C....
seq04 C..TG..T.. .C.....A.. G......C.. TT.C......
seq05 CA.TG..T.. .C....TA.. G......CC. TT.C......
seq06 C...G..... CC..T..A.A ....AA.C.. TT..G.....
seq07 CA.TG..... CC..TC.A.A ...CA..CC. TT...C....
seq08 CA.TG...T. GCT....A.. G..C..TC.. T.......T.
seq09 CA.TG..... GC.C..T... ...CA..C.A T.........
seq10 CA.TG...T. .C.....A.. G..C...C.. T....C....
```

**Answer *5.10.*** For the 20 polymorphic sites, the frequencies of the "minor allele" (the rarer of the two at each site) are 2, 7, 2, 6, 2, 2, 7, 7, 2, 5, 1, 1, 2, 1, 1, 7, 3, 1, 2, and 1. The folded observed spectrum is

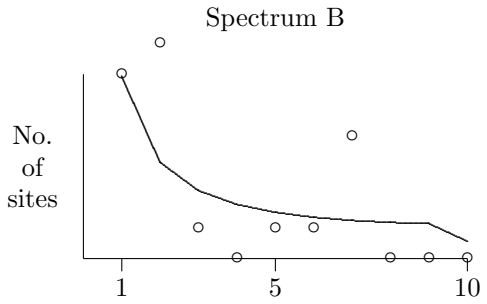
| Minor allele count | Observed spectrum |
|---|---|
| 1 | 6 |
| 2 | 7 |
| 3 | 1 |
| 4 | 0 |
| 5 | 1 |
| 6 | 1 |
| 7 | 4 |
| 8 | 0 |
| 9 | 0 |
| 10 | 0 |

**Answer *5.12.*** To estimate $\theta$, note that $S = 20$, so $\hat{\theta} = 20/\sum_{i=1}^{K-1} 1/i = 5.63739$. The expected folded spectrum is

Data set B

| Minor allele count | Expected spectrum |
|---|---|
| 1 | 5.93410 |
| 2 | 3.13188 |
| 3 | 2.21074 |
| 4 | 1.76169 |
| 5 | 1.50330 |
| 6 | 1.34224 |
| 7 | 1.23899 |
| 8 | 1.17446 |
| 9 | 1.13887 |
| 10 | 0.56374 |

**Answer *5.14.***




The graph shows the observed spectrum as open circles and the expected one as a solid line.

**Answer *6.2.*** The values below indicate that relative error is small when $p$ is less than about 0.1.

| $p$ | relerr |
|---|---|
| 0.30 | 0.22 |
| 0.20 | 0.14 |
| 0.10 | 0.07 |
| 0.08 | 0.05 |
| 0.06 | 0.04 |
| 0.05 | 0.03 |

**Answer *6.4.*** As with the Jukes-Cantor model, we begin with

$$\begin{aligned} q &= e^{-2\lambda t} + (1 - e^{-2\lambda t})/2 \\ &= \frac{1}{2} + \frac{1}{2} e^{-2\lambda t} \end{aligned}$$

When at least one perturbation has occurred, the sites are identical with probability 1/2. This accounts for the value "1/2" that appears above. Now substitute $2u = \lambda$ (because only half of perturbations are mutations), $p = 1 - q$, and $K = 2ut$:

$$p = \frac{1}{2} - \frac{1}{2} e^{-2K}$$

Solving for $K$ gives

$$K = -\frac{1}{2}\log_e(1-2p)$$

which is analogous to Eqn 6.1. This was introduced in 1919 and is called "Haldane's mapping function" [3]. It has often been used to make linkage maps of chromosomes.

**Answer *6.6.*** (a) For the three species, Jukes-Cantor yields the following estimates of $K$:

| | $K$ estimates | | |
|---|---|---|---|
| | BN Rat | Mouse | Human |
| Brown Norway Rat | — | 0.38 | 0.20 |
| Mouse | | — | 0.40 |
| Human | | | — |

(b) The mouse-human distance is *much* larger than the rat-human distance. (We hope students will contemplate the causes of this difference, but grades will not be based on this issue.)

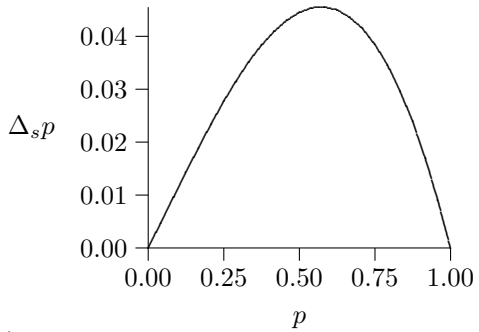
**Answer *7.2.*** (a) The mean relative fitness is $\bar{w} = 1 - 2pqhs - q^2s = 0.966$. (b) The marginal fitness of $A_1$ is $w_1 = p + q(1-hs) = 0.996$. That of $A_2$ is $w_2 = p(1-hs) + q(1-s) = 0.959$. (c) In the following generation, the expected frequency $A_1$ is $p' = (p^2w_{11} + pqw_{12})/\bar{w}$, as shown on Gillespie's p. 62. This gives 0.2061.

**Answer *7.4.*** (a) Gillespie makes his fitnesses relative to that of genotype $A_1A_1$. To obtain these relative fitnesses, divide each absolute fitness by $W_{11}$. This gives relative fitnesses $w_{11} = 1$, $w_{12} = 0.769$, and $w_{22} = 0.692$. In Gillespie's fitness scheme, $w_{12} = 1-hs$, and $w_{22} = 1-s$. This implies that $s = 1 - w_{22} = 0.308$, and $h = (1 - w_{12})/s = 0.75$.

(b) Gillespie's Eqns. 3.2–3.3 express $\Delta_s p$ in terms of $p$, $q = 1-p$, $s$, $h$, and $\bar{w} = 1 - pqhs - q^2s$. Note that $q$ and $\bar{w}$ depend on $p$ and must therefore be recalculated for each value of $p$ that you graph. I did these calculations using a Python script. Here are a few of the results in tabular form:

| $p$ | $\Delta_s p$ |
|---|---|
| 0.0000 | 0.0000 |
| 0.0417 | 0.0047 |
| 0.0833 | 0.0095 |
| ... | ... |
| 0.9167 | 0.0170 |
| 0.9583 | 0.0090 |
| 1.0000 | 0.0000 |

The graph of these results looks like this:



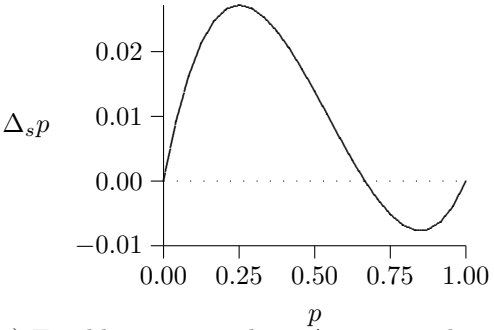

(c) Equilibria occur where $\Delta_s p = 0$. As the graph shows, the only such points are at $p = 0$ and $p = 1$. Only the second of these is stable. This is because $\Delta_s p > 0$ whenever $0 < p < 1$. Consequently, $p$ always moves *toward* the equilibrium at $p = 1$ but *away from* the one at $p = 0$.

**Answer *7.6.*** (a) Gillespie makes his fitnesses relative to that of genotype $A_1A_1$. To obtain these relative fitnesses, divide each absolute fitness by $W_{11}$. This gives relative fitnesses $w_{11} = 1$, $w_{12} = 1.111$, and $w_{22} = 0.889$. In Gillespie's fitness scheme, $w_{12} = 1-hs$, and $w_{22} = 1-s$. This implies that $s = 1-w_{22} = 0.111$, and $h = (1-w_{12})/s = -1$.

(b) Gillespie's Eqns. 3.2–3.3 express $\Delta_s p$ in terms of $p$, $q = 1-p$, $s$, $h$, and $\bar{w} = 1 - pqhs - q^2s$. Note that $q$ and $\bar{w}$ depend on $p$ and must therefore be recalculated for each value of $p$ that you graph. I did these calculations using a Python script. Here are a few of the results in tabular form:

| $p$ | $\Delta_s p$ |
|---|---|
| 0.0000 | 0.0000 |
| 0.0417 | 0.0092 |
| 0.0833 | 0.0162 |
| ... | ... |
| 0.9167 | –0.0063 |
| 0.9583 | –0.0039 |
| 1.0000 | –0.0000 |

The graph of these results looks like this:



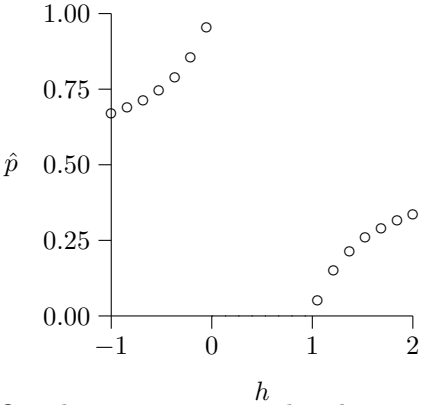

(c) Equilibria occur where $\Delta_s p = 0$. There are three such points: at $p = 0$, at $p = 1$, and at $p = 2/3$. $\Delta_s p$ is positive to the left of the intermediate equilibrium but negative to the right. Consequently, the intermediate equilibrium is stable and

the two extreme equilibria are unstable.

**Answer *7.8.*** No calculation is required. Because the two homozygotes have equal fitness and the heterozygote is superior, the stable equilibrium frequency of $A_1$ is 1/2.

**Answer *7.10.*** $\hat{p}$, the equilibrium value of $p$, is equal to $(h-1)/(2h-1)$ (see Gillespie's Eqn. 3.4). Here are a few values and then the graph:

| $h$ | $\hat{p}$ |
|---|---|
| –1.0000 | 0.6667 |
| –0.8421 | 0.6863 |
| –0.6842 | 0.7111 |
| ... | ... |
| 1.6842 | 0.2889 |
| 1.8421 | 0.3137 |
| 2.0000 | 0.3333 |




(a) Overdominance occurs when $h < 0$ and generates a stable equilibrium within the interval $(0, 1)$. (b) Incomplete dominance occurs when $0 < h < 1$. In this case, selection is directional, and there are no internal equilibria. (c) Underdominance occurs when $h > 1$ and generates an unstable internal equilibrium. The only stable equilibria are at $p = 0$ and $p = 1$.

**Answer *7.12.*** *(a)* In the first generation, the frequency of white flowers is $P_{rr} = 200/1000 = 0.2$. At Hardy-Weinberg equilibrium, this frequency is $q^2$, where $q$ is the frequency of allele $r$. This implies that $q = \sqrt{0.2} = 0.447$. The frequency of allele $R$ is therefore $p = 1 - q = 0.553$. *(b)* In the 2nd generation, $P_{rr} = 250/1000 = 0.25$, so $q' = \sqrt{0.25} = 0.5$, and $p' = 1 - q' = 0.5$. *(c)* Because $w_{RR} = w_{Rr} = 1$, Eqn. 7.1 simplifies to $p' = p/(1 - q^2 s)$, or $s = (1 - p/p')/q^2$. With our data, this is $s = (1 - 0.553/0.5)/0.2 = -0.53$.

**Answer *7.14.*** Because the number of new advantageous mutants is proportional to population size, but the fixation probability for such mutants is essentially independent of population size. The rate of adaptive evolution is thus proportional to population size.

Some students may want to answer this question more formally, so here is the formal version: let $u$ represent the mutation rate for advantageous alleles whose fitness in heterozygotes is $1 + s/2$, relative to that of the wild-type allele. The number of such mutations in the population as a whole is $2Nu$, and each of them has probability $s$ of ultimate fixation. For this category of adaptive mutations, the expected number fixed per generation is therefore $2Nus$, an increasing function of population size.

**Answer *7.16.*** The substitution rate per million years is $10^6 \times 2Nu$ times the answers from the preceding question. (a) $9.07 \times 10^{-11}$ if $N = 10000$; or (b) $6.89 \times 10^{-4}$ if $N = 1000$; or (c) $1.99 \times 10^{-3}$ if $N = 100$.

**Answer *7.18.*** The substitution rate per million years is $10^6 \times 2Nu$ times the results from the `pfix` function (Gillespie's Eqn. 3.22) defined above. (a) 0.044 if $N = 10000$; or (b) 0.0051 if $N = 1000$; or (c) 0.0024 if $N = 100$.

**Answer *8.2.*** (a) The gamete frequencies are $x_1 = 80/165 = 0.485$, $x_2 = 30/165 = 0.182$, $x_3 = 10/165 = 0.061$, and $x_4 = 45/165 = 0.273$. (b) The allele frequencies at the two loci are $p_A = x_1 + x_2 = 0.667$, and $p_B = x_1 + x_3 = 0.545$. (c) $D = x_1 x_4 - x_2 x_3 = 0.121$. (d) $r^2 = D^2/[p_A(1 - p_A)p_G(1 - p_G)] = 0.267$.

**Answer *8.4.*** (a) $x_1 = 0.1$, $x_2 = 0.3$, $x_3 = 0.5$, and $x_4 = 0.1$. (b) $p_A = x_1 + x_2 = 0.4$, and $p_G = x_1 + x_3 = 0.6$. (c) $D = x_1 x_4 - x_2 x_3 = -0.14$. (d) $r^2 = D^2/[p_A(1 - p_A)p_G(1 - p_G)] = 0.3403$.

**Answer *8.6.*** (a) After 1 generation, $D = 0.2997$. (b) After 100 generations, $D = 0.2714$.

**Answer *8.8.*** One approach is to add the two expressions on the right side. If both equations are correct, then this sum should equal $D - D = 0$. Summing the two equations gives

$$\begin{aligned} D - D &= x_1 - p_A p_B + x_2 - p_A(1 - p_B) \\ &= x_1 + x_2 - p_A(p_B + 1 - p_B) \\ &= x_1 + x_2 - p_A \\ &= p_A - p_A \\ &= 0 \end{aligned}$$

The two expressions sum to zero, so the two equations are equivalent.

**Answer *8.10.*** The gametic fitnesses are $(w_1, w_2, w_3, w_4) = (1 + s, 1 + s, 1, 1)$, and mean fitness is $\bar{w} = \sum_i w_i x_i = 1 + s/2$.

**Answer *8.12.*** (a) Diversity should be lower at nearby loci. (b) This effect is larger in regions of low recombination.

**Answer *8.14.*** (a) The approximate formula gives $c = 10^{-8} \times 5 \times 10^7 = 0.5$. (b) Haldane's mapping function gives $c = (1 - \exp(2 \times 10^{-8} \times 5 \times 10^7))/2 = 0.316$. (c) The approximation is poor when the distance between loci is large.

**Answer *8.16.*** We are looking for sites such that $c/s < 0.1$, or $c < 0.1s$, where $c$ is the recombination rate between the site in question and the selected site. For this problem, $s = 0.001$, so $c < 10^{-4}$. We established above that when $c$ less than 0.1 or so, it is approximately equal to $\lambda k$, where $\lambda = 10^{-8}$, and $k$ is the distance between the sites in base pairs. Our largest $c$ value is much smaller than 0.1, so we can use this approximate formula. Using this formula, $c = 10^{-8}k < 10^{-4}$, so $k < 10,000$. Variance is removed from a region twice this size, because the region extends for 10,000 bases each way from the selected site. The size of the affected region is therefore 20,000 bases, or 20 kb.

**Answer *8.18.*** $s = 0.05$.

**Answer *8.20.*** There are many possible answers. Here are a few examples. (a) Sites 1–12 in sequences 41–42. (b) Sites 1–15 in sequence 31. (c) Sites 38-42 in sequences 40–49, excluding sequence 46.

**Answer *9.2.*** The data tell us that $P_{11} = 200/500 = 0.4$, and $P_{12} = 100/500 = 0.2$, so the frequencies of alleles $A_1$ and $A_2$ are $p = P_{11} + P_{12}/2 = 0.5$, and $q = 1 - p = 0.5$. To calculate $F$, we might work with the formula for any of the three genotype frequencies. Let's use $P_{12} = 2pq(1 - F)$. Rearranging this gives $F = 1 - P_{12}/2pq$, which works out to equal 0.6.

**Answer *9.4.*** $P_{11} = 0.313$

**Answer *9.6.*** 1/16

**Answer *9.8.*** $P_{11} = 0.19$, $P_{12} = 0.42$, and $P_{22} = 0.39$.

**Answer *9.10.*** $P_{11} = p^2 + pqf$, $P_{12} = 2pq(1 - f)$, and $P_{22} = q^2 + pqf$.

**Answer *10.2.*** We will work with equation 10.1, which is a function of $H_S$ and $H_T$. Let us begin with the formula for $H_S$:

$$\begin{aligned} H_S &= \sum c_i 2p_i(1 - p_i) \\ &= 2\sum c_i p_i - 2\sum c_i p_i^2 \\ &= 2p - 2\sum c_i p_i^2 \end{aligned}$$

The sum here is a weighted average over subpopulations, so think of it as an expectation, and use the hint mentioned in the text of this question: $\sum c_i p_i^2 = V + p^2$, where $V$ is the variance of the $p_i$. This gives

$$H_S = 2p - 2p^2 - 2V = 2p(1 - p) - 2V \qquad \text{(A.1)}$$

By definition, $H_T = 2p(1 - p)$. Substitute this, along with equation A.1, into equation 10.1, and you will get

$$\begin{aligned} F_{ST} &= (H_T - H_S)/H_T \\ &= \frac{2p(1-p) - 2p(1-p) + 2V}{2p(1-p)} \\ &= V/p(1-p) \end{aligned}$$

**Answer *10.4.*** The question only asks for $F_{ST}$, but I'm providing more detail: $H_S = 0.185367$, $H_T = 0.482516$, and $F_{ST} = 0.615832$.

**Answer *10.6.*** $P_{AA} = 0.125$, $P_{Aa} = 0.25$, and $P_{aa} = 0.625$,

**Answer *10.8.*** It implies that $Nm = 2$.

**Answer *12.2.*** Line 2: $\bar{X} = 4.48$; $V = 0.959$.

**Answer *12.4.*** Line 4: $\bar{X} = 3.55$; $V = 0.876$.

**Answer *12.6.*** Variance between lines: 0.1454.

**Answer *12.8.*** Between plus within: $0.1454 + 0.8998 = 1.04520$. In question 12.5, I got $V_P = 1.002$, which is pretty close.

**Answer *12.10.*** Broad-sense heritability.

**Answer *12.12.*** $V_A = h^2 V_P = 0.05 \times 1.002 = 0.0501$. The dominance variance is $V_D = V_G - V_A = 0.1454 - 0.0501 = 0.0953$. Most of the genetic variance seems to be dominance variance. This is disappointing because selection responds only to additive variance.

**Answer *12.14.*** The response to selection is $R = h^2 S$, as shown in Gillespie's Eqn. 6.11. If you made $h^2$ larger while keeping $S$ the same, the response would be larger. But this might be hard to do, because if each plant is represented by its average seed weight, differences among plants will probably be small. For this reason, it might be necessary to reduce $S$ in order to make selection feasible.

**Answer *12.16.*** The correlation between parent and offspring is $h^2/2$. Thus, we estimate $h^2$ as twice the observed correlation. For span, this gives $h^2 = 0.904$.

**Answer *12.18.*** The change per generation is $0.157 \times 28/10 = 0.4396$ in. If this represents a response to selection, then it should equal $V_A\beta$. The next step is to convert our estimate that $h^2 \approx 1$ into an estimate of $V_A$. We know that $h^2 = V_A/V_P$ (Gillespie's Eqn. 6.4), and we know that $V_P = 6.26$ (table 12.2 of this homework). This implies that

$V_A \approx 6.26$ and that the selection gradient is $\beta \approx 0.4396/6.26 = 0.07$. In words, this says that an extra inch of stature implies nearly a 10% increase in fitness—*very* strong selection. It seems implausible that small differences in stature could have such large effects on fitness. It is more likely that the observed trend reflects changes in the environment.

# Bibliography


[1] Ronald A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[2] R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, 2004.

[3] J. B. S. Haldane. The combination of linkage values, and the calculation of distance between loci of linked factors. *Journal of Genetics*, 8: 299–309, 1919.

[4] Clifford J Jolly and Fred Plog. *Physical Anthropology and Archeology*. Alfred A. Knopf, fourth edition, 1986.

[5] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[6] Newton E Morton, Shirley Yee, DE Harris, and Ruth Lew. Bioassay of kinship. *Theoretical Population Biology*, 2(4):507–524, 1971.

[7] Masatoshi Nei and Sudhir Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, 2000.

[8] Karl Pearson and Alice Lee. On the laws of inheritance in man. I. inheritance of physical characters. *Biometrika*, 2(4):357–462, November 1903.

[9] Alan R. Rogers. Lecture notes on gene genealogies. URL `http://content.csbs.utah.edu/~rogers/ant5221/ggeneal.pdf`. Unpublished manuscript, 2013.

[10] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Menlo Park, 1977.

[11] Ziheng Yang. *Computational Molecular Evolution*. Oxford University Press, Oxford, 2006. ISBN 0198567022.