Study Guide for Human Evolutionary Genetics

Alan R. Rogers

January 8, 2023

1 Python

Here is an example of the sort of question that we plan to ask:

1. Write down the output that would be produced by the following snippet of Python code.

```
s = 0
for i in range(3):
    s += i*i

print(s)
```

We would expect to see something like this as an answer:

```
Program prints "5".
```

You only need to know about Python constructs that have been used in the course, either in JEPy or in one of the labs. We went through the code in those projects and made a series of short snippets like the one above. Here are a few more, just to give you the flavor:

2.
```
print("a %s b %d:%3.1f" % ("hey",12,1.2))
```

3.
```
def f(x, y):
    return x*y

print(f(3,2))
```

4.
```
x = 5

print("%d %s" % (x, x*"*"))
```

5. Here's one that often comes in handy when doing calculations with gene genealogies:

```
print(sum([1/i for i in range(1,3)]))
```

   This last snippet is the easy way to calculate $\sum_{i=1}^{2} 1/i$.

6.
```
from random import random
samp_size = 100000000
n = 0
for i in range(samp_size):
    if random() < 0.7:
        n += 1

p = n/samp_size
print(p)
```

The last snippet above is the longest one in our list.

2 Probability

7. Make a table showing the relative frequencies of the values in these data: [A, A, T, A, G].

8. In his urn experiment, Kerrich drew two balls from an urn *without* replacement. Imagine a version of experiment in which each trial begins with 5 red balls and 3 black ones. What is the probability that, in a single trial, both of the balls drawn are red?

9. Here is the probability distribution of two variables, X and Y:

X	Y	$\Pr[X,Y]$
0	0	0.2
0	1	0.2
1	0	0.1
1	1	0.5

   What are $E[X]$, $E[Y]$, $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, $\mathrm{Cov}[X,Y]$, $E[XY]$, and $E[XY^2]$.

10. Be able to manipulate the properties of expectations listed at the bottom of the first column of page 6, in JEPr. We'll ask you for things like $E[4X]$, $E[Y/2]$, and $E[X+Y]$. You should be able to get such answers quickly, using the expectations you calculated in the previous item.

11. An urn contains 40 black balls and 60 red balls. You choose a ball at random, put it back, and

choose another at random. What are the possible outcomes and their probabilities?

12. In a toss of two fair dice, one red and one black, what is the probability of observing a red 4 *or* a black 5?

13. You toss a fair coin three times, receive $1 for each head and nothing for tails. Let X represent the total number of dollars you receive on all three tosses. What is the probability distribution of X? (In other words, what are the possible values of X and their probabilities.)

14. Suppose that, in a class of 40 students, 20 are women. If we choose a student at random from the class, what is the probability that this student is a woman?

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

16. Now you select 3 students at random *with* replacement. The number of men in this sample is a random variable, which may equal 0, 1, 2, or 3. What is the probability distribution of this random variable? (You may answer either by listing the probability of each outcome or by writing down a formula. Don't bother with the calculator.)

17. JEPr gives three formulas for the variance: $E[(X - E(X))^2]$, $E[X^2] - E[X]^2$, and $E[X(X - E[X])]$. Be able to calculate a variance using all three.

18. The preceding item is about the variance as an expected value. But we also calculate variances from data, and in that case the variance is a statistic rather than an expected value. These are distinct quantities, even though we use the same word for them. Make sure you understand the difference between a statistic and an expected value, and make sure you can calculate either sort of variance.

19. JEPr also gives three formulas for the covariance: $E[(X - E(X))(Y - E[Y])]$, $E[XY] - E[X]E[Y]$, and $E[X(Y - E[Y])]$. Be familiar with these too, both as statistics and as expected values.

20. JEPr discussed the following probability distributions: (1) binomial, (2) Bernoulli, (3) Poisson, (4) uniform, (5) exponential, and (6) normal. You're expected to know the mean, variance, and distribution function of each distribution.

21. Under what cirumstances would each distribution be plausible? For example, which distribution would you use to model the number of clicks emitted by a Geiger counter during some fixed interval of time? Suppose you spin a bottlecap 30 times and count the number of times it lands "concave-side-up." Which distribution would best model that? Which distribution would make sense as a model of the length of a coalescent interval? (This last question won't make sense until you've read the chapter on Gene Genealogies in Rogers's lecture notes.)

22. For which distribution is the mean equal to the variance? For which is the mean equal to the standard deviation?

3 Alleles, loci, frequencies, and heterozygosity

23. Discuss the two inconsistent sets of definitions of "locus," "gene," and "allele."

24. A locus with two alleles is called "biallelic." Data from such a locus may be in the form of a list of genotypes like this:

    ```
    ['AA', 'AA', 'AT', 'TA', 'TT']
    ```

    or in the form of genotype counts like this:

    ```
    AA: 2, AT: 2, TT: 1
    ```

    Given either type of data, you should be able to calculate (1) genotype frequencies, (2) allele frequencies, and (3) genotype frequencies expected at Hardy-Weinberg equilibrium.

25. Define genetic drift. What events in the lives of real organisms contribute to drift?

26. Describe the Urn Model. What do the balls in the urn represent? What do the balls drawn from the urn represent? How does the urn model behave?

27. Under the influence of drift alone, heterozygosity ($\mathcal{H}$) declines according to

    $$\mathcal{H}_t = \mathcal{H}_0(1 - 1/2N)^t$$

    Here are a few examples of questions that we might ask about this formula:

    (a) We might ask you to complete some step or steps of the derivation, so be sure you know how it goes.

(b) Describe it in plain English. What do the symbols $\mathcal{H}_t$, N, and t mean?

(c) Genetic drift has its largest effect in small populations. How is this fact reflected in the formula?

28. At equilibrium between mutation and genetic drift, and assuming the infinite alleles model of mutation, the expected heterozygosity is

$$\mathcal{H} = \frac{4N\mu}{4N\mu + 1}$$

There are several kinds of questions that we might ask about this formula. For example:

(a) What model of mutation does this formula assume?

(b) We might ask you to complete some step or steps of the derivation, so be sure you know how it goes.

(c) What is the distinction between this formula and the formula $H = 2p(1-p)$ for heterozygosity at Hardy-Weinberg equilibrium? Both formulas predict heterozygosity at equilibrium, so why are they different? (Hint: What is meant by "equilibrium" in the two contexts?)

(d) What is genetic drift, and how does it affect variation within and among populations?

(e) Ditto for mutation?

(f) How are these two effects captured by the equation above?

(g) The formula above predicts the *average* heterozygosity. Values at individual loci vary widely at any given time, from 0 up through levels of heterozygosity far in excess of the average. Why?

(h) Suppose that (a) the average heterozygosity equals 0.1, and (b) that $\mu = 10^{-6}$. What can you conclude about N? (Assume that the assumptions that went into the formula above are all correct.)

4 Gene genealogies, spectrum, and mismatch distribution

The questions that follow concern n copies of a neutral DNA sequence. These sequences were drawn from a population of constant size, and we are interested in their gene genealogy.

29. How many coalescent intervals are in this gene genealogy?

30. What is the hazard of a coalescent event within the interval with i lineages?

31. What is the expected length of this interval in generations, assuming that the diploid population size is N?

32. What is the expected age (in generations) of the last common ancestor?

33. What is the expected total branch length of the tree? (In other words, the expected sum of the lengths of all branches.)

34. What is the expected number of segregating sites, assuming that there are L sites, and that the mutation rate is u per site per generation?

35. What are the expected values of π, the mean pairwise difference per site, and Π, the mean pairwise difference per sequence, using the same assumptions?

Finally, some general questions:

36. We have introduced a variety of statistics for measuring genetic variation: H, π, Π, and S. How does population size affect the expected values of these quantities?

37. *Why* does population size affect measures of genetic variation? (We are expecting prose here, not formulas.)

38. Coalescent theory predicts that, if one population is twice as large as another, it should have much more heterozygosity. Is real data consistent with this prediction? Why or why not?

39. The expected number, S, of segregating sites (Eqn. 5.4 of *Lecture Notes on Gene Genealogies*) is proportional to population size. Thus, a population twice as large should have twice as many segregating sites, given samples of equal size. What if the populations were of equal size, but one sample was twice as large as the other? Would the expected number of segregating sites double if the sample size K were twice as large? (Hint: calculate $\sum_{i=1}^{K-1} 1/i$ for $K = 2, 3, 4, 5$. Here's how: `sum([1.i for i in range(1,6)])`.)

40. Theory predicts that the values of π and $S/\sum_{i=1}^{K-1} 1/i$ should be similar, because both of these estimate the same parameter. Yet this was not true of the human data that we discussed in class. What might account for this discrepancy?

41. Why are coalescent intervals longer near the root of the tree? (Answer in words, not in formulas.)

42. Why do large populations have more genetic variation than small ones? (Answer in words, not in formulas.)

43. How are the sizes and shapes of gene genealogies affected when the population size, N, increases? (Assume selective neutrality.)

44. Be able to calculate the site frequency spectrum and the mismatch distribution from data (for small, simple data sets).

45. How do they respond to changes of population size (as in the previous section)?

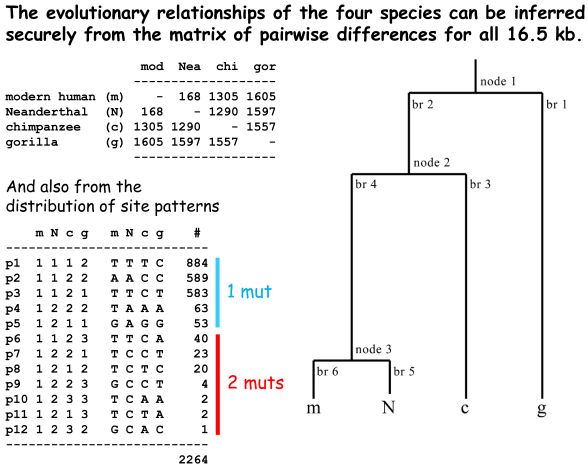
5 Neutral theory of molecular evolution

The "neutral theory" is a set of ideas that explain why inferred rates of molecular evolution tend to be similar for a given gene or gene region, especially in closely related species. To illustrate this pattern we considered the mitochondrial phylogeny of a modern human, a Neanderthal, a chimpanzee and a gorilla, as summarized in this slide from the neutral-theory lecture.

The tree was drawn by a method that allows the branch lengths to vary with the number of nucleotide substitutions (mutations) that are inferred to have occurred on each branch. The tip of the gorilla branch (br 1 + br 2, g) was aligned arbitrarily with the tip of the modern human branch (br 4 + br 6, m). The chimp branch (br 3, c) is not constrained by that aesthetic choice, yet its tip aligns very closely with those of the modern and Neanderthal humans, showing that its mitochondrial genome has evolved at a very similar rate, following separation of the chimp and human lineages around 6 million years ago.

In the table of site patterns, the nucleotide state in the modern human sequence (m) is arbitrarily represented as "1", whether it's G, A, T or C. The other allele(s) are represented as "2" and if necessary "3". A particular instance of that class of patterns is then shown in the second block of columns, and the total number of that class of patterns in the last column. Those sum to 2264 sites that vary in the 4-sequence alignment; that's roughly 14%, implying that 86% don't vary.

46. Why is the first pattern (P1) the most common?



47. Why are p4 and p5 the least frequent of the patterns requiring just one mutation? Is their difference (63 versus 53 instances) likely to be significant (statistically and/or scientifically)? Why or why not?

48. Why do we say that the number of new mutations per generation at a given site in a given population is $2N$?? What is the meaning (definition and/or units) of each of the three terms in this product?

49. What assumptions are we making when we say that the chance of ultimate fixation for any such new mutation is $1/2N$?

50. How does the chance of ultimate fixation change for mutations with deleterious effects on the organism's phenotype, and hence its fitness? For mutations with beneficial effects?

51. For the triplet codons within a protein-coding sequence, at which of the three positions (1st, 2nd, 3rd) is a new mutation most likely to be effectively neutral? And at which of the three positions are mutations least likely to be effectively neutral? Please explain your predictions.

52. This question and the next one make use of unpublished mitochondrial genome sequences for six individual right whales. The 13 protein-coding genes taken together contain a total of 11,379 nucleotide positions (3,793 codons). 280 codons are variable, meaning that at least one of the whales has at least one nucleotide that differs in state from the others at that position. There are 55 variable first positions, 13 variable second positions, and 218 variable third positions. (These numbers add up to more than 280

because six of the codons vary at two different positions—either first and second, or first and third.) Is this pattern consistent with neutral-theory expectations? What does it imply (quantitatively and/or qualitatively) about the way natural selection really does act on typical mutations at the three codon positions?

53. The six right whales represent three distinct geographic populations that are thought to have separated roughly 2-3 million years ago (Mya): the southern right whale (Eubalaena australis), the North Atlantic right whale (E. glacialis), and the North Pacific right whale (E. japonica). Their closest living relative is the bowhead whale (Balaena mysticetus, formerly called the "Greenland right whale"), from which their ancestral lineage separated around 6 Mya. Use the pairwise difference matrices below to infer the mitochondrial phylogeny for all seven whales, using the simple clustering method called UPGMA (https://en.wikipedia.org/wiki/UPGMA). First make the tree implied by the differences at third codon positions only, and then make the tree implied by the differences at first and second positions only.

In which tree do you have more confidence, and why? Note that the UPGMA method assumes a "clocklike" accumulation of differences with time. The two kinds of sites clearly have very different clock rates, but if we recalibrate by assuming that the separation between the bowhead and right-whale lineages occurred around 6 Mya, do both trees give similar dates for the splits among the three right-whale species?

Would you expect to get the best result by combining all of the data (that is, by adding the two distance matrices)? Feel free to give that a try. In any case, please explain your reasoning.

Raw distance matrix, position 3

	Bmy	Eau1	Eau2	Egl1	Egl2	Eja1	Eja2
Bmy	0	452	458	450	453	467	467
Eau1	452	0	36	108	113	130	130
Eau2	458	36	0	104	109	134	132
Egl1	450	108	104	0	9	130	128
Egl2	453	113	109	9	0	133	131
Eja1	467	130	134	130	133	0	34
Eja2	467	130	132	128	131	34	0

Raw distance matrix, positions 1–2

	Bmy	Eau1	Eau2	Egl1	Egl2	Eja1	Eja2
Bmy	0	115	118	132	134	119	120
Eau1	115	0	13	35	35	26	27
Eau2	118	13	0	40	40	31	30
Egl1	132	35	40	0	6	45	46
Egl2	134	35	40	6	0	45	46
Eja1	119	26	31	45	45	0	13
Eja2	120	27	30	46	46	13	0

6 Selection

54. Assuming fixed, constant fitnesses for the three genotypes at a biallelic diploid locus, what is the *mean fitness* of the population?

55. What are the *marginal fitnesses* of the two alleles?

56. What is next generation's expected allele frequency (p' or q'), as a function of these quantities?

57. How can an observed rate of allele-frequency change be used to estimate the relative fitnesses of two alleles (for example, s, if we know the value of h)? This was the subject of some homework problems.

58. Selection is slow when a recessive allele is rare. Why?

59. Selection is slow when a dominant allele is common. Why?

60. A rare allele tends to spread if its heterozygote is fitter than the common homozygote. Why? Why doesn't the fitness of the rare homozygote matter?

61. What stable equilibria exist if $w_{11} > w_{12} > w_{22}$?

62. What stable equilibria exist if $w_{11} < w_{12} > w_{22}$?

63. When the heterozygote has highest fitness, and the two homozygotes differ in fitness, which allele is most common at the stable equilibrium?

64. What stable equilibria exist if $w_{11} > w_{12} < w_{22}$?

65. Selection pushes allele frequency in the direction that increases the mean fitness of the population. Consequently, stable equilibria occur at peaks in the graph of mean fitness against allele frequency.

66. How long does it take for an advantageous allele to sweep from frequency 0.01 to 0.99 in a large population, assuming that allelic effects are additive (no dominance) and that the fitness of the favored homozygote is 1.018 times that of the other homozygote?

Answer: The coefficient of selection is $s = 0.018$, so the answer is $18/s = 1000$ generations.

7 Interactions of mutation, drift and selection

67. What is the fixation probability for a newly arisen neutral mutation?

68. What is the *approximate* fixation probability for a newly arisen mutation that is favored by a selection coefficient of s (relative fitness $1 + s$) in the homozygous state, and $s/2$ in the heterozygous state? Does the population size N affect our answer? If so, how?

69. What is the expected rate at which adaptive mutations will fix within a species, given a certain population size (N), selective advantage (s), and genome-wide rate of mutation (U) to alleles with advantages of about that size?

70. Why are most deleterious alleles (the ones segregating at appreciable frequencies) at least partly recessive?

71. Under what conditions will a deleterious (harmful) mutation have nearly as good a chance of fixing as a neutral mutation?

72. Do we expect more adaptive evolution in large or small populations? Explain.

73. Do we expect to find more or fewer harmful mutations segregating in a selfing plant species (for example, rice) than in a predominantly outbreeding relative? Explain.

8 Multiple loci

74. Be familiar with Gillespie's notation (x_1, x_2, x_3, and x_4) for frequencies of two-locus haplotypes, where each locus has two alleles. Remember that $p_A = x_1 + x_2$ and $p_B = x_2 + x_4$. (Sec. 4.1, p. 102 of Gillespie.) Also, $D = x_1x_4 - x_2x_3$.

75. If we gave you the frequencies of these four gamete types, you should be able to give us the allele frequencies and D (the coefficient of linkage disequilibrium). Or the other way around.

76. What are the relationships among these two measures of linkage disequilibrium: D and r_H? (Here, r_H refers to the gametic correlation between two loci. We used this symbol in the lab manual. Gillespie uses "r" for the same purpose.)

77. Locus B has two alleles: B_1 (with frequency p_1), and B_2 (with frequency $1-p_1$). A new mutation arises at a linked locus, A. What is the probability that this mutant chromosome carries allele B_1?

78. Mammalian genomes (including ours) are typically about 3 Gbp (billion base pairs) in length, but most of that DNA is non-functional "junk" that doesn't code for proteins or structural RNAs, and doesn't play any important role in the regulation of gene expression. Some of these neutral nucleotides are close to functional genes, and others are far away. How would you expect proximity to functional genes to affect the variability of these neutral nucleotides? Are the ones near genes more likely to be polymorphic, or less?

79. When an allele sweeps to fixation at locus A, (a) what happens to the frequencies of neutral alleles at linked loci? (b) What happens to D?

80. Linkage disequilibrium is often obvious in tables of DNA sequence (or SNP) data. For examples, see our slides. Be able to recognize it when you see it.

81. What is genetic draft? What problem was it invented to explain? How is it affected by the following parameters: N, μ, c (recombination rate)?

9 Inbreeding

82. What is inbreeding depression, and what causes it?

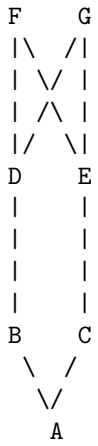
83. The inbreeding coefficient (F) can be interpreted as a measure of either (a) departure from Hardy-Weinberg genotype frequencies, (b) inbreeding within a pedigree, (c) inbreeding between random individuals within a population, or (d) the effect of genetic drift. In the first sense, it is the proportional reduction in heterozygosity relative to that expected at Hardy-Weinberg equilibrium. In the second sense, it is the probability of identity by descent (IBD) relative to the oldest generation in the pedigree. Senses 3–4 are like 2 but refer to a random individual within a population rather than to a specific individual.

84. Be able to express the frequencies of the three genotypes at a biallelic locus (P_{11}, P_{12}, and P_{22}) in terms of p_1 and F.

85. The formulas in the preceding question can be used to represent the effect of non-random mating in a single generation, or the effect of drift across many generations. How do the interpretations of p and F change in these two situations? (Hint: your answer should involve the "reference generation.")

86. What is the connection between inbreeding and genetic drift?

87. Consider the following genealogy:



In this genealogy, A is the offspring of B and C, which are the offspring of D and E, and so on. (a) The two genes that united to form A may be identical by descent (IBD) from which ancestors? (b) For each of these ancestors, calculate the probability of IBD. (In other words, calculate the contribution of each loop in the pedigree to A's inbreeding coefficient. (c) What is A's inbreeding coefficient? (d) What is the coefficient of kinship of D and C? (e) In what way, if any, do we expect A to differ *genetically* from typical (outbred) members of the population? Explain briefly. (f) In what way(s) might we expect A to differ *phenotypically* from outbred members of the population? Explain briefly.

Answers: (a) F, and G; (b) 1/32 for F or G; (c) $F = 2 \times 1/32 = 1/16$; (d) $f_{DC} = 1/8$. (e) Individual A will be heterozygous at fewer loci than will the average member of the population, because there is an appreciable probability that the gene copies he inherited from his two parents were identical by descent from a common ancestor (F or G). (f) Individual A is likely to be somewhat shorter than the average individual and somewhat less healthy, because he will have fewer loci that exhibit heterozygote advantage and more that are homozygous for recessive deleterious alleles.

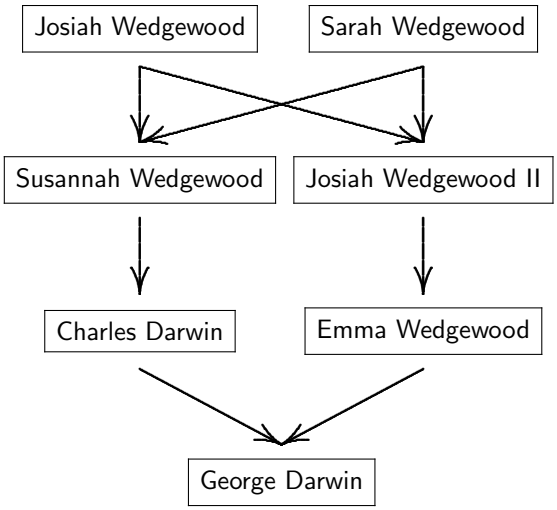


Figure 1: Genealogy of George Darwin

88. Figure 1 shows the genealogy of George Darwin, one of Charles Darwin's sons. What is his inbreeding coefficient?

10 Non-random mating

89. What is F_{ST}, and what does it measure? How does it increase with time among a group of isolated populations, such as those in Buri's experiment? What equilibrium does it reach if the effects of drift are opposed by gene flow?

90. 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? **Answer:**

$$1/9 = \frac{1}{4Nm+1} \Rightarrow Nm = 2$$

11 Human evolutionary history

91. What is known about the history of human population size during the past two million years? When was the population large? When was it small? How do these events relate to (a) the

expansion of *Homo erectus* out of Africa, about 2 Ma ago, (b) the expansion of modern humans out of Africa about 50 ka ago, (c) the coldest part of the last ice age about 20 ka ago, and (d) the beginning of the Holocene (the warm period in which we now live) about 12 ky ago?[1]

92. When and where did Neanderthals live? Which modern populations carry DNA derived from Neanderthal ancestors? In modern Eurasians, what fraction of the DNA derives from Neanderthals?

93. Expect the same question about Denisovans, except that we don't know much about when and where they lived. (They seem to have been contemporaneous with Neanderthals, but probably lived in Asia, Indonesia, and/or New Guinea.)

94. At some loci, archaic alleles seem to have been favored by selection in modern populations. In which functional category (or categories) of gene is this beneficial effect strongest? Why might archaic alleles have been beneficial to modern humans at these loci?

95. At other loci, archaic alleles were selected against in modern humans. How do we know? Why might we expect deleterious alleles to have been common in archaic genomes?

96. What do we know about the "neandersovan" ancestors of Neanderthals and Denisovans? When did they separate from the lineage leading to modern humans? When did they split into Neanderthals and Denisovans? How large was their population? With whom did it interbreed?

Some vocabulary: The "mesolithic" of Europe refers to the period after the Ice Ages but before agriculture. During this period, Europeans lived by gathering and hunting medium-sized animals. The following period, the "neolithic," Europeans farmed and raised domestic animals but did not yet make metal tools. After the neolithic came the ages of copper, of bronze, and of iron.

97. What do we know about the history of skin and eye color in Europe? How did these characters change from the mesolithic (say 8000 bp) to the neolithic (say 6000 bp) to the bronze age (say 4000 bp) to the recent period?

[1]To prepare for this question, look at the slide labeled "PSMC estimates from autosomes" from the lecture "Population history from whole genomes." (You can also find this graph in the article by Li and Durbin.) What does the graph say about population size at the times listed in the question?

98. Same question, but about lactase persistence: the ability to digest milk sugar (lactose) beyond childhood.

12 Quantitative traits

99. What are quantitative traits?

100. Why are they often normally (or near-normally) distributed?

101. What is the phenotypic variance (V_P)?

102. What are the relationships (mathematical and biological) among the variance components V_P, V_G, V_A, V_D and V_E?

103. What is the narrow-sense heritability (h^2) and why is it important?

104. How can each of the variance components be estimated for real populations?

105. Why do heritability estimates for one population have little validity for another population of the same species? (In other words, why do we have to make the estimates anew, from actual data, to be confident of their values in a given population?)

106. Related question: if a trait is highly heritable within each of two populations, why does this *not* imply that the difference between those populations is genetic?

107. Response to selection can be predicted either from

$$\begin{aligned} R &= h^2 S, \qquad \text{or from} \\ \Delta x &= V_a \beta \end{aligned}$$

Gillespie discusses only the first formulation. The second is covered by Rogers [3], which is available on the class website. Be familiar with both forumulations and able to manipulate them.

108. What evidence suggests that long-term rates of quantitative trait evolution are not limited by amounts of heritable genetic variation?

References

[1] M. Hasegawa, H. Kishino, and T. Yano. "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". *Journal of Molecular Evolution* 22 (1985), pp. 160–174.

[2] T. H. Jukes and C. R. Cantor. "Evolution of protein molecules". In: *Mammalian Protein Metabolism*. Ed. by H. N. Munro. New York: Academic Press, 1969, pp. 21–132.

[3] Alan R. Rogers. *Response to Natural Selection on a Quantitative Character*. 2022. URL: https://content.csbs.utah.edu/~rogers/tch/ant5221/unprotected/Rogers-lande.pdf.