OEB 242
Midterm Review Practice Problems – Answer Key

(1) **Loci, Alleles, Genotypes, Haplotypes**

(a) Define each of these terms.

Locus *– a genetic site*
Allele *– a genetic variant, particular to a given genetic locus, segregating in a population*
Genotype *– the combination of alleles observed or postulated at a given locus in a given individual in a population*
Haplotype *– the combination of alleles across two or more loci observed or postulated in an individual in a population*

(b) We used the expression $\binom{n}{k}$, which is equal to $\frac{n!}{k!(n-k)!}$ and represents sampling *without replacement* where order doesn't matter, to calculate the number of heterozygotes possible in a diploid population in PS2 #1. The formula for sampling *with replacement* where order doesn't matter is: $\binom{n+k-1}{k}$. Use this information to give the number of possible genotypes in an octoploid species at a locus with 11 alleles. How many of these genotypes are heterozygous (in the sense that *at least one* of the eight alleles an individual carries differs from the other 7)?

In these expressions, k *represents the ploidy and* n *represents the number of segregating alleles (how many ways are there to put* n *objects into* k *slots)? So,* $\binom{11+8-1}{8} = \frac{18!}{10!\,8!}$ = 43758*. There are 11 possible 'true homozygotes' (i.e. individuals with all 8 of their alleles matching each other) so there are* 43758 – 11 = 43747 *possible heterozygotes.*

(c) Suppose we add to our analysis another locus with *j* alleles. How many haplotypes are possible between the two sites?

Given two loci, the number of possible haplotypes is simply the product of the number of variants at the two sites. (The number of possible genotypes at either site is not relevant.) So, there are 11*j *possible haplotypes.*

(2) **Hardy-Weinberg and χ^2**

(a) Suppose we have a diploid population in HWE. Define the allele and genotype frequencies at a biallelic site.

Given alleles A and a:

P(A) = p
P(a) = q = 1-p
P(AA) = p^2
P(Aa) = 2pq
P(aa) = q^2

(b) Suppose we remove half of the heterozygotes from the population. What will the new allele frequencies be (in terms of their frequencies in the first generation, p_0 and q_0)?

If our population is of size N, then we have reduced our heterozygote count from 2pqN individuals to pqN individuals. Our population size is now N – pqN. We must take both of these things into account:

$$P(A) = p_1 = \frac{2 * p_0^2 * N + p_0 q_0 * N}{2 * (1 - pq) N}$$

$$P(a) = q_1 = \frac{2 * q_0^2 * N + p_0 q_0 * N}{2 * (1 - pq) N}$$

Note that we can't simply transform the allele frequencies, we have to take account of the population size, even though we removed an equal number of A alleles and a alleles.

(c) How long will it take the population to reach a new equilibrium? What will the HWE values be (in terms of their frequencies in the first generation, p_0 and q_0)?

Under the Hardy-Weinberg model where mating is truly random (with imagined infinite population size, no selection, etc. etc.) HWE only takes one generation to reach. So, having disturbed the population, it will reach a new equilibrium after one more round of random mating. We can simply plug our expressions for p_1 and q_1 into the Hardy-Weinberg equations to get the new equilibrium values:

$$P(AA) = p_1^2 = \left[\frac{2 * p_0^2 * N + p_0 q_0 * N}{2 * (1 - pqN)}\right]^2$$

$$P(Aa) = 2p_1q_1 = 2 * \frac{2 * p_0^2 * N + p_0 q_0 * N}{2 * (1 - pqN)} * \frac{2 * q_0^2 * N + p_0 q_0 * N}{2 * (1 - pqN)}$$

$P(aa) = q_1^2 = [\frac{2 * q_0^2 * N + p_0 q_0 * N}{2 * (1 - pqN)}]^2$

(d) Attempt to perform and interpret a test for HWE at a biallelic site in a diploid population with the following genotype frequencies: P(AA) = .65, P(Aa) = .3, P(aa) = .05. (What information is missing that prevents you from doing so? Why do you need this information? Try a few different values and see how it affects your interpretation of your test.)

Get allele frequencies from genotype frequencies:
P(A) = p = .8
P(a) = q = .2

Get expected genotype frequencies using HWE:
$p^2 = .64$
$2pq = .32$
$q^2 = .04$

We can't apply χ^2 *because we need to have observed and expected values for genotype counts; frequencies do not suffice. After all, in a small sample just a few individuals with unexpected genotypes can easily happen by chance, offsetting our genotype frequencies, whereas in a large sample the same proportional increase in individuals seems less likely under the null hypothesis.*

Moreover, we know that the χ^2 *distribution with* k *degrees of freedom represents the sum of squares of* k *independent standard normal random variables – we don't expect the permissible deviation in frequency to be standard normal; but we can impose this assumption on the deviation in* count *of however many independent classes we have.*

If N = 100:

	AA	***Aa***	***aa***
obs	*65*	*30*	*5*
exp	*64*	*32*	*4*

$\chi^2 = .390625$
df = (3 classes) – (1 for fixing N) – (1 for estimating p) = 1
p = .53 (we cannot reject the null hypothesis of neutral HWE).

If N = 10000:

	AA	***Aa***	***aa***
obs	*6500*	*3000*	*500*
exp	*6400*	*3200*	*400*

χ^2 = 39.0625
df still equals 1
Now, p is infinitesimally small – essentially zero, and we can reject H_0.

(3) **Linkage Disequilibrium**

(a) For two biallelic loci (A/a and B/b):
Recall that we define the linkage disequilibrium parameter, D, such that P(AB) = pApB + D. It represents the deviation from expected haplotype frequencies assuming linkage equilibrium (i.e. complete independence of sites)
Show that D = pAB*pab – paB*pAb.

P(AB) = pApB + D
P(ab) = papb + D (by analogy)
P(aB) = papB – D (because pa = 1-pA; pb = 1-pB)
P(Ab) = pApb – D

We can verify the given equation by plugging in these expressions:

*D = [(pApB + D) * (papb + D)] – [(papB – D) * (pApb – D)]*

D = [pApBpapb + D papb + D * pApB + D^2] –*
*[papBpApb – D * papB – D * pApb+ D^2]*

D = (D papb + D * pApB) –*
*(D * papB – D * pApb)*

D = D[papb +pApB + papB + pApb]

D = D[(1-pA)(1-pB) + pApB + (1-pA)pB + pA(1-pB)]

D = D, QUED ;P

(b) Calculate and interpret D and D' if we have P(AB) = .5, p(ab) = .05, p(aB) = .25, p(Ab) = .2. Do we need to know the sample size to make this calculation?

D = (.5)(.05) – (.25)(.2) = -.025
D' = D/Dmin given that D is negative
Dmin = min[-(.7.75), -(.3*.25)] = -.075*
D' = -.025/-.075 = .3333

Thus, D is a third of its most extreme value. The negative sign simply reflects which allele we have chosen to define as big-A or big-B.

No, these statistics are defined to be independent of sample size. Note, however, that sample size is often worth considering when making sense of LD: in small samples, rare variants tend to get lost, which often makes LD appear more robust than it is.

(c) Calculate r^2. How is this information different from the information we get when we calculate D'?

Remember that r^2 *is a measure of correlation – it is not the same r that we used to represent the frequency of recombination (which determines the rate of LD decay).*

$r^2 = D^2/(pApBpapB)$
=(-.*025*)^2/(.7*.3*.75*.25)
= .01587

The square root of which is .125, indicating a weak correlation between the two sites. Thus, these mutations happened with enough time in between them that they don't carry a lot of mutual information. If on the other hand one of the alleles (say a*) represented a mutation that had happened much earlier in the same lineage where the* b *mutation had just occurred, then we could have different haplotype frequencies that would result in a higher value of* r^2 *without necessarily having the same magnitude of effect on D'. This is a subtle point; see textbook pp. 84-85 for an example.*

(4) Drift: The Wright-Fisher and Moran Models

(a) Another way to think of drift in a diploid population has some features of both the Wright-Fisher and Moran models. In this third model, we have 2N timesteps per generation (a la Moran). At each timestep, we randomly sample one allele with replacement from the population and put it in our new generation (a la Wright-Fisher).

On this model, what is the probability that a particular allele (i.e. only one copy exists in the population) has at least one copy in the next generation? Assume an infinite population size and give an exact answer.
(Hint: $\lim_{\epsilon \to 0}(1+\epsilon x)^{1/\epsilon} = e^x$)

The probability that this allele doesn't get chosen to reproduce at a given timestep is $1 - \frac{1}{2N}$*. Since each timestep is independent, the*

probability of this happening for a whole generation is $(1 - \frac{1}{2N})^{2N}$.
Using the hint given, we can see that for a theoretically infinite population, this value, the probability of extinction, becomes 1/e. So, the probability of non-extinction is 1 – 1/e.

(5) **The Coalescent**

(a) What is the expected time to the first coalescent event for a sample of 10 alleles in a population of 100? In a population of size 1000?

Using the Kingman coalescent,
$E[T_{10}] = 4N/k(k-1) = 400/(90) = 4.444$ generations for N = 100
$= 4000/(90) = 44.444$ generations for N = 1000

(b) Which of the two values is the more certain?

Again using the Kingman coalescent, the variance on our estimate is $\frac{16N^2}{(k(k-1))^2}$ *. Thus, a larger N gives us larger variance on our estimate of the time to coalescence, and we have more confidence in our estimate when N = 100 than when N = 1000.*

(6) **Mutation**

(a) Show that, under the infinite-sites model, the expected number of segregating sites in a sample of *n* chromosomes in a diploid population is $\theta * \sum_{i=1}^{n-1} \frac{1}{i}$.

Under the infinite-sites model, each mutation happens at a new locus and creates a new allele. So, we can think of the expected number of segregating sites in terms of mutations distributed randomly on a coalescent tree: sum up all the branch lengths (in units of generations, but will be functions of the population size) and multiply by the per-generation mutation rate, U (analogous to the per-generation, per-site rate, μ).

We can use the Kingman coalescent to do this. Recall that T_i represents the expected time to the first coalescent event <u>from</u> a sample of size i. Thus, our tree will always have i branches of length T_i that lead up to each coalescent event, for i between n (at the leaves of the tree) and 2 (coalescing at the root).

$$E(total\ length\ of\ tree) = E(\sum_{i=2}^{n} i * T_i)$$

$$= \sum_{i=2}^{n} i * E(T_i)$$

$$= \sum_{i=2}^{n} i * \frac{4N}{i(i-1)}$$

$$= 4N * \sum_{i=2}^{n} \frac{1}{(i-1)}$$

$$= 4N * \sum_{i=1}^{n-1} \frac{1}{i}$$

$$E(S) = U * E(total\ length\ of\ tree)$$

$$= 4NU * \sum_{i=1}^{n-1} \frac{1}{i}$$

$$= \theta * \sum_{i=1}^{n-1} \frac{1}{i}$$

(b) Given the following sample of 6 chromosomes genotyped at 6 variable sites out of a stretch of 50bp, calculate Π. Use this information to estimate θ two ways. How would you use these values if you wanted to calculate Tajima's D? What would the sign of Tajima's D be, and how would you interpret it here? What is the equilibrium homozygosity under the infinite alleles model for each of these estimates of θ? (Note that these estimates may disagree dramatically given our small and artificial sample.)

A	G	T	A	T	T
A	G	T	C	G	A
A	G	C	C	T	A
G	G	C	C	T	T
G	G	C	C	G	T
G	T	C	C	G	T

Π represents per-site pairwise diversity: if you picked two random chromosomes and compared them at a random site, what's the probability that they would differ? So, we need to count how many mismatches there are and how many possible mismatches there are.

Each site makes the following contribution:

9	*5*	*8*	*5*	*9*	*8*	*mismatches*

for a total of 44. We are making $\binom{6}{2}$ *= 15 chromosome comparisons in our sample set. So we have 44/15 = 2.9333 differences per sequence on average. (Pick a few random pairs of chromosomes and count how many differences there are to convince yourself that this makes sense).*

To get a per-site value, we divide by L = 50bp. We get Π = .05866. *To estimate* θ, *we must remember two predictions of the infinite-sites model:*

$E(\Pi) = \theta$ *(because E[*T_2*] is 2N, thus 4N generations of evolutionary time separate a given pair)*

$E(S) = \theta * \sum_{i=1}^{n-1} \frac{1}{i}$ *(as we showed in part a)*

Thus, our two estimates are $\widehat{\theta_\pi}$ *and* $\widehat{\theta_S}$*. They are, respectively, .05866 and .0525 (using an exact value for the sum rather than the ln(n) approximation).*

Tajima's D takes (and then normalizes) the difference $\widehat{\theta_\pi}$ - $\widehat{\theta_S}$*, which here would be positive. This indicates a surplus of intermediate frequency alleles, which contribute more to the per-pairwise comparison number of mismatches than do rare frequency alleles.*

If we could extrapolate from this small sample to the population, we would infer shallow coalescence times for this locus (such that most mutations occur on shared branches of our coalescent tree). Such a tree structure would indicate, for example, a decrease in population size or balancing selection.

Homozygosity is the probability that two randomly sampled alleles at a locus are different. Under the infinite alleles model, we can recursively define homozygosity as follows:

$F_t = (1-\mu)^2(\frac{1}{2N}) + (1-\frac{1}{2N})(1-\mu)^2 F_{t-1}$

Where the left hand term represents the probability that the two alleles coalesce in the preceding generation and neither have mutated, and the right hand term represents the probability that although they do not coalesce in the preceding generation, they were homozygous before then, and furthermore have not mutated. We set $F_t = F_{t-1}$ *and treat* μ *and* μ^2 *as approximately zero to find the equilibrium homozygosity value:* $\frac{1}{1+\theta}$ *. This represents the scenario where mutation creates diversity at the same rate that drift removes it.*

$\widehat{F_\pi}$ = .94459
$\widehat{F_S}$ = .95011

The difference between these answers makes sense for our sample: since half of our sites are singletons, estimating $\hat{F}$ based on per-site heterozygosity will lead to a larger value (you're more likely to find a match if you sample two random alleles) than if we were to estimate it based on the number of segregating sites.