Note: All data and examples in this exam are fictitious.

1. (25 pts) A researcher samples pink salmon from a river and measures allele frequencies at a protein coding locus. Two alleles can be detected using gel electrophoresis (the protein products of the two alleles differ in electrical charge). The following genotypes were observed in adult fish:

Genotype	Count	H-W Expectations
BB	1332	1216.8
Bb	456	686.4
bb	212	96.8
Total	2000	

(a) (3 pts) Fill in the H-W expectations in the table above.

See above.

(b) (6 pts) Give three possible explanations for the discrepancy between these observations and H-W. (You can assume it is statistically significant.)

Acceptable answers include subpopulation structure (or migration; these are essentially the same explanation here), non-random mating, presence of an undetected null allele, selection against the heterozygote (underdominance), or linkage disequilibrium with a selected locus.

Overdominance would cause too many heterozygotes, but there are too few. Varying population size does not disrupt H-W. Random chance is very unlikely to cause such a large discrepancy, given the big sample size. I gave only partial credit to answers that said "linkage disequilibrium" because by itself, disequilibrium does not disrupt H-W, but in the presence of a linked selected locus it could.

(c) (8 pts) If 20% of the fish in the sample originate in a hatchery, and the frequency of B in hatchery fish is 0.3, what genotype frequencies would we expect? Could this level of gene flow explain the discrepancy from H-W? (You do not need to do the statistical test; just make an informed guess.) Show your work.

Suppose the hatchery fish were in H-W. There are 400 such fish, and with pB=0.3 they would have the H-W expectations shown below. Subtract them from the total fish and we have a remaining population with pB=(1296+288/2)/1600=0.9. The final column below shows H-W for a population with that pB. The match is in fact unreasonably good. (If I were reviewing a paper with these data I would seriously wonder whether, rather than estimating the 20% hatchery fish with pB=0.3 from independent data, the authors had picked one or the other number to optimize the match to expectations....)

Genotype	*Hatchery*	*Remainder*	*Remainder expectation*
BB	*36*	*1296*	*1296*
Bb	*168*	*288*	*288*
bb	*196*	*16*	*16*
Total	*400*	*1600*	*1600*

Problem 1 continued

The researcher finds a river which has no hatcheries. He takes genetic samples from newborn fish at the spawning grounds, and then takes samples from the adults two years later when they return to spawn (pink salmon spawn exactly two years after birth). (Note that the newborn and adult fish are samples from the same population, but are not necessarily the same fish.)

Genotype	Newborn	Adult	Fitness (unnormalized)	Fitness (normalized)
BB	723	734	1.015	1.125
Bb	256	231	0.902	1.0
bb	21	45	2.143	2.375
Total	1000	1000		

(d) (3 pts) Fill in the relative fitnesses in the table above.

See above.

(e) (5 pts) Given these fitnesses and the current frequency of B, if we allow the system to evolve undisturbed for a long time, what frequency of the B allele do we expect? Assume that the population size is large and mutation rate is small, so mutation and drift can be disregarded. Show your work.

We need the equilibrium t/(s+t). From the fitnesses above, s=-0.125 and t=-1.375, so the equilibrium pB=0.9166. If B is more frequent than this, it will fix; if it is less frequent, it will be lost. The observed pB is either 0.851 (in the newborns) or 0.849 (in the adults). Either way it is below the tipover point and b will fix. We can verify this by noting that pB got smaller in one generation of selection, moving in the direction of being lost.

2. (27 pts) A particular region of the hamster genome is transcribed at a high level. It contains a potential coding sequence which aligns to a functional mouse gene called KUH. However, hamster KUH has a predicted protein structure which is quite different from KUH in mouse or other rodents. We consider two hypotheses:

   (H1) Hamster KUH is a real gene, but it has adapted to a new function.

   (H2) Hamster KUH is a non-functional pseudogene.

   We count the number of sites in the KUH coding sequence at which a synonymous or nonsynonymous mutation would be possible. We then count the number of positions at which hamster KUH has a synonymous or nonsynonymous difference from mouse KUH, obtaining the following data:

	Synonymous	Non-synonymous
Possible differences	78	215
Actual differences	19	52

   (a) (3 pts) What is the dN/dS ratio for these data?

   dN=52/215, dS=19/78, so dN/dS is 0.9929.

   (b) (6 pts) Is it more supportive of H1 or H2? Explain briefly. (If you were unable to calculate dN/dS, explain what each possible outcome would mean.)

   This is very close to 1, meaning that the gene does not seem to care whether changes are synonymous or non-synonymous; this is typical of pseudogenes and unusual for a functional gene. It is possible, however, that the observed ratio is an average between a conserved region with dN/dS<1 and a region selected for diversity or rapid divergence with dN/dS>1.

   Using humans as an outgroup, we measure polymorphism within the hamster population and divergence between hamster and mouse for this gene, and for a large intron, believed to be neutral, in an unlinked gene. Note that the intron and the KUH sequence are not the same length. We obtain the following data:

	Intron	KUH
Fixed differences between mouse and hamster	114	45
Polymorphic sites within hamster	38	16

   (c) (6 pts) Do these data give more support to H1 or to H2? Explain briefly.

   These are the inputs for an HKA test, but in lectures I didn't really specify which one is the numerator and which is the denominator! In any case, the ratio of polymorphism to divergence is pretty similar between the intron and KUH, suggesting that they are evolving in about the same way. As we assume the intron is non-selected, that implies that KUH is also non-selected. However, lacking a statistical test here, I accepted any reasonable interpretation of the numbers obtained. (They look a lot more different if you use polymorphism as the numerator.)

Problem 2 continued

(d) (8 pts) A colleague who prefers experimental solutions to statistical ones uses CRISPR to delete the hamster KUH gene. She obtains four homozygous deletion hamsters and successfully raises them to adulthood. She argues that this settles the question: H2 (pseudogene) must be correct. Critique this argument.

It is a terrible argument, but surprisingly common....

Showing that animals survive to adulthood in the lab does not show that they (a) are fertile, (b) would survive in a natural environment, (c) would be competitive with normal members of their species. An example of this is that an early screen for "all functional genes in Drosophila" missed the key circadian-rhythm locus, which is important to flies in the wild but irrelevant when they are raised under constant light conditions. Remember that a tiny difference in fitness is enough for selection to work with in a large population.

Showing that KUH- animals have no selective disadvantage would be extremely laborious, requiring hundreds or thousands of animals and many years; doing it in a natural environment is particularly hard. I know of only one experiment of this magnitude in mammals: Barrett et al. (2019) Science. Worth looking at!

3. (29 pts) We are studying a single-celled, diploid eukaryote which can reproduce either sexually (meiosis, recombination, mating) or asexually (mitosis).

We start with a large wild population of the organism, which can be assumed to be in Hardy-Weinberg equilibrium and linkage equilibrium, and to be well mixed. We use two large random samples of this population to start two separate laboratory populations. These are maintained at identical large, constant population sizes in identical environments. The only difference is that one reproduces sexually and the other reproduces asexually (we label them the Sexual and Asexual populations). Assume that the treatment used to make them reproduce in a particular way has no side effects.

After many generations, we survey the two populations. For each question below, briefly explain your answer. ("They will be the same" is a legitimate option.)

(a) (4 pts) Which population will have more linkage disequilibrium between loci along a chromosome?

Asexual, because recombination in the Sexual lineage will tend to break up LD.

(b) (5 pts) Which population will have a shorter average time back to the population common ancestor of its alleles?

Asexual. There are two reasons. First, even though they are nominally diploid, there are really only N competing lineages in the Asexual population, so the effective population size is lower. Second, whereas genetic draft and hitchhiking will reduce N_e close to the selected locus in Sexual, they will reduce it everywhere in Asexual, leading to a more recent TMRCA.

This is a subtle point and I gave partial credit for the other answer.

(c) (5 pts) If the two populations were put in competition, which would you expect to win? If it could go either way, what factors are particularly important?

On the whole I'd expect Sexual to do better. It can reassort two good mutations into the same individual; it will have a wider variety of haplotypes available, which should speed up adaptation; it won't suffer from Muller's Ratchet. Asexual might win if sex were costly or risky (phages?), there were overdominant loci which it could fix, as shown in the next subproblem, or there were massive co-adapted gene complexes such that recombination led to bad haplotypes.

Problem 3 continued

The wild population is polymorphic at a locus called *pwl*; we measure the frequency of the + allele as 0.57. In our lab environment, the two homozygotes (+/+ and *pwl*/*pwl*) each have a fitness of 0.9, and the heterozygote (+/*pwl*) has a fitness of 1.0.

(d) (4 pts) After a long time, what frequencies of the three **genotypes** (measured prior to the action of selection) do you expect in the Sexual population?

This is symmetrical overdominance. The allele frequencies will go to 0.5 and the genotype frequencies are then 0.25, 0.5, 0.25.

Several students went wrong by giving H-W for this generation, ignoring "After a long time".

(e) (4 pts) ... in the Asexual population?

The Asexual population will simply fix the best genotype, +/pwl. They can do this because there is no reassortment from sex, so the three genotypes behave more like three alleles, and this is the best one.

(f) (6 pts) After a long time, which population will end up with higher mean fitness *for this locus*? Explain briefly. You do not need to calculate the mean fitness.

Asexual, because it does not create inferior homozygotes through sexual reassortment. Its relative fitness will go to 1, whereas Sexual will go to 0.95 (I did not require this to be calculated).

4. (20 pts) A region of the human X chromosome is analyzed in a large European sample and found to have substantially fewer variable sites than the European whole-genome average. The researcher attributes this to a selective sweep in the European population.

   State, and briefly explain, two alternative hypotheses for this finding.

   (a) (10 pts) Hypothesis 1:

   (b) (10 pts) Hypothesis 2:

   Reasonable hypotheses include:

   The X has a lower population size than an autosome because males have only one copy, so we expect shorter coalescence time and lower variable sites.

   The X might tolerate bad recessive mutations more poorly due to being hemizygous in males, leading to quicker removal of these mutations and thus fewer variable sites. (X inactivation, which can cause a phenotype in heterozygous females, could make this effect even stronger.)

   This region by chance has a more recent common ancestor than average.

   This region has a lower mutation rate than average.

   There is less recombination on the X (because when it is in a male it can't recombine) and this intensifies hitchhiking and genetic draft, leading to shorter TMRCA and lower variation.

   This region has more essential loci than most, so stronger selection.

   There was a recent selective sweep in this region.

   Less reasonable ideas include:

   Introgression of this region (should make it more variable, not less, unless the introgressed version was the only one still present–and that would still need explaining).

   Several answers which seemed to be trying to explain having more LD or lower pairwise differences between individuals, but those were not mentioned in the problem. For example, if there were more epistasis on the X, this would lead to higher LD, but why would it lead to lower variable sites?

$2N$	Generations to common ancestor of 2 lineages
$4N$	Approximate generations to common ancestor of population
$p^2 + 2pq + q^2$	Hardy-Weinberg
$1/2N_e = \sum_g 1/2N(g)$	Population size varying over time
$N_e = 4N_f N_m/(N_f + N_m)$	Unequal sex ratio
$F \approx 1/(1 + 4N_e\mu)$	Proportion of homozygotes with drift and mutation
$pA = \nu/(\mu + \nu)$	Mutation equilibrium
$pA = t/(s + t)$	Overdominant or underdominant equilibrium
$2s$	Survival probability of new mutant with multiplicative fitness
$\bar{w} = pAAwAA + pAawAa + paawaa$	Mean population fitness
$D = pAB - pApB$	Disequilibrium coefficient
$D' = D/max(D)$	Normalized disequilibrium coefficient
$r^2 = \frac{D^2}{pA\ pa\ pB\ pb}$	Squared disequilibrium correlation coefficient
$D_n = (1 - c)^n D_0$	Decay of disequilibrium
$D = \theta_\pi - \theta_S$	Tajima's D; θ_π is mean difference, θ_S is count of SNPs
Approximate breakpoints:	
$4N_e\mu$	Drift vs. mutation
$4N_e s$	Drift vs. selection
$4N_e m$	Drift vs. migration
$4N_e c$	Drift vs. recombination
m/s	Migration vs. selection