Short Problems:

1. (This problem is inspired by Khan et al. 2011 but the specific numbers are fictitious, and probably much bigger than realistic.) We start with a standard lab strain of haploid *E. coli* which we will call "wildtype". We isolate two mutants at different loci (loci *pykF* and *topA*) and measure their fitness in all possible combinations. Assume that these two loci are the only loci that vary in our experiment. In the following table, all-caps indicates the wildtype allele.

Genotype	Relative fitness
PYKF TOPA	1.0
pykF TOPA	1.0
PYKF topA	1.1
pykF topA	1.4

   (a) (1 pt) If fitnesses were multiplicative, what fitness would you predict for the double mutant pykF/topA? *1.1*

   (b) (1 pt) What kind of epistasis is this? (Positive, negative, sign?) *Positive. It does not change the direction of selection but does change the magnitude.*

   (c) (1 pt) If we wanted to fix the advantageous *topA* mutation (via selection) as quickly as possible, would we prefer to start with *PYKF* wildtype cells or *pykF* mutant cells? Explain briefly. *Mutant. The proportional advantage of topA is greater in a pykF background, so it will fix faster.*

2. Consider two loci G and H in (diploid!) humans. We use pedigrees to identify haplotypes (combinations of alleles that are inherited together) in Europeans, and obtain the following data:

Haplotype	Number observed
G H	37
G h	43
g H	3
g h	17
Total	100

   (a) (2 pts) Calculate the disequilibrium coefficient D for these data. *pGH (observed) - pGpH (expected) = .37 - .32 = 0.05*

   (b) (2 pts) What is the maximum D for these allele frequencies? (Hint: genotype frequencies must not be < 0 or > 1.) *For positive D this is the lesser of pgH and pGh, which is 0.08.*

   (c) (1 pt) Calculate D' = D/max(D) *0.05/0.08 = 0.625*

   (d) (4 pts) The usual assumption, when we see a situation like this, is that G and H are linked on the same chromosome. Describe at least two situations in which we might see disequilibrium even if G and H were on different chromosomes, and explain each one briefly. *Our population could really be a mix of two populations with different allele frequencies. There could be strong selection against particular G/H combinations (epistasis). Our organism might reproduce mostly or completely by cloning, so that even separate chromosomes behave as if linked. We haven't covered this yet, but intense inbreeding such as self-fertilization could also have this effect.*

3. (3 pts) In general, what allele frequencies at two loci allow the largest possible disequilibrium D between them? Explain briefly. *Both loci should have pA=pa=0.5. This gives the biggest possible difference between what haplotype frequencies we expect (0.25) and the maximum we can observe (0.5).*

4. We have a chip which can genotype 1 million common polymorphisms (SNPs) in Europeans, and use it to do genome-wide association studies (GWAS) in which we search for SNPs that are strongly correlated with a phenotype of interest, such as a disease.

   (a) (3 pts) A very common finding is that the SNP with the highest association with the phenotype is not the closest SNP to the mutation that actually causes the phenotype, but is a little way off. Why might this be? *The strength of association is based not only on distance to the causal variant, but on the strength of LD, which depends on the allele frequencies and the historical accident of what haplotype the causal mutation fell on. A further-away variant*

with allele frequencies that perfectly match the causal variant may be much more strongly associated than a closer variant which is much more common than the causal variant.

Another way of looking at this is that for two variants that fall within the same local coalescent tree in the coalescent-with-recombination, the highest correlation will come if they both arose on the same branch of the tree, and if they arose on unrelated branches the correlation may be quite poor.

Several people misinterpreted this question as asking how a non-coding SNP could be strongly associated with a trait; I gave one point for correct answers to this question but it is not what I asked.

(b) (2 pts) Why do we prefer a European chip and not a global chip when doing GWAS in Europeans? *Just about everyone pointed out that different sites are SNPs in different populations and that one doesn't want to pay to genotype sites that don't vary. I was also thinking that LD patterns vary between populations, but on consideration that's a reason not to use non-European test subjects, not a reason not to use a global chip.*

Long Problem:

Modern European, Asian, and Oceanian humans apparently have some ancestry from the archaic Neanderthal and Denisovan populations, whereas modern Africans do not.

The $FOXP2$ gene lies at the center of a large "archaic ancestry desert" where no modern humans appear to have Denisovan or Neanderthal alleles. Mutation in $FOXP2$ in a modern human family caused pervasive language and speech disorders.

We can suggest at least two different hypotheses for the cause of this "desert."

H1: Some time after the intermating of humans with Neanderthals and Denisovans, a favorable mutation arose in a human haplotype in this region and swept to fixation, removing any Neanderthal or Denisovan alleles that might previously have been present in the population.

H2: Prior to intermating, humans had already fixed a favorable allele in some locus in this region; thus, Neanderthal and Denisovan haplotypes were eliminated shortly after being introduced, as the human haplotypes were superior.

1. (2 pts) Can you suggest a way to distinguish H1 from H2, given a big sample of sequenced human genomes? In other words, what might differ in a modern human sample depending on whether H1 or H2 is correct? *Several good ideas were proposed. In H1 there is a single, recent favored haplotype so it should have little variation; in H2 the favored haplotypes are much older and should vary more. H1 also predicts a recent sweep in Eurasians only; if African haplotypes are quite different here (more variables, different alleles) that would support H1. I would add that seeing the same desert for both Neanderthals and Denisovans, who were well separated and presumably had different harmful alleles, weakly suggests H1.*

2. (2 pts) The human and Neanderthal $FOXP2$ genes code for the same amino acid sequence (differing by two amino acids from the chimp sequence). Why doesn't this refute the idea that $FOXP2$ might be the locus responsible for the desert? *Because there can be selection on promoters, enhancers, or even silent-site variants, and this can matter just as much as selection on coding variants.*

3. (4 pts) There are a number of deserts in the modern human genome, but this is by far the biggest. List at least two factors that could influence the size of a desert. Explain each briefly. *Lower local recombination rate, perhaps due to a lack of hot spots, makes a bigger desert as the favored haplotype breaks down more slowly. Stronger selection makes a bigger desert as the favored haplotype fixes faster and there is less time for recombination. Multiple selected loci in the same area could fuse into a big desert, especially if they are epistatic. Randomness of the coalescent could make this area particularly young, hence with particularly long haplotypes.*

4. (2 pts) It has been suggested that modern African populations intermated with a different archaic population, but no DNA-bearing fossils of this hypothetical ancestor are available. Explain briefly why it is so much harder to be sure about archaic ancestry when the archaic sequence is unavailable. What other phenomena could be mistaken for archaic

ancestry? *When ancient DNA is available, we can detect not only that sequences are unusual for modern humans, but that they match the archaic sequence. Without this, we could mistake regions that are randomly older than usual, under balancing selection, or have a high local mutation rate for regions introgressed from an archaic population. Frustrating!*