Solution key - 2010 7.012 Problem Set 5

Question 1

A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single base pair in the genome differs between members of a species or paired chromosomes in an individual. By convention this base pair change is represented as one nucleotide — A, T, C, or G — of the base pair.

a) Circle the correct option from the following choices. The SNPs may exist …

i. only within the coding sequences of genes.
ii. only within the non-coding regions of genes.
iii. only in the coding and non-coding regions of the genes.
[✓ iv. in the coding or non-coding regions of the genes or in the intergenic regions between the genes.]

b) In which region or regions (*coding sequence of the gene/ non-coding sequence of the gene/ intergenic sequence*) would you expect the SNP to be if …

i. it changes the amino acid sequence of the protein which is produced? Include **all** the possible options**.**

The SNP can only be in the coding sequence (i.e. exons) of the gene that codes for the amino acids of the protein.

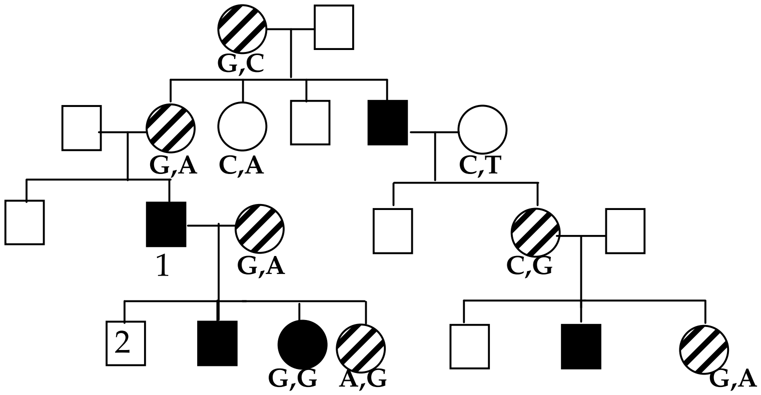
ii. it does not change the amino acid sequence of the protein which is produced? Include **all** the possible options and give an **explanation** for each selected option**.**

The SNP may be in the coding region assuming that it creates a silent mutation that does not change the amino acid sequence of the protein. It may also be in the non-coding sequence (i.e. introns) of the genes or the intergenic regions between the genes because these segments do not code for the amino acids of the protein.

c) You are studying a family with an inherited disease associated with Gene Z. You identify two SNPs that are linked to Gene Z, SNP1 is closer to Gene Z than is SNP2. Explain why you would use SNP1 **and not SNP2** as a marker to follow the inheritance pattern of this disease in the family.

You want to use SNP 1 as a marker because the closer the two loci are the lesser is the chance of a recombination event occurring between them.

d) Below is the pedigree of a family with this disease. All the individuals that show the disease phenotype are shaded and the carriers are striped. Also listed are the alleles of a SNP *(A, G, T, C)* for some individuals. *Note: you may assume that this SNP is tightly linked to Gene Z and may be used as a marker for the disease. Assume complete penetrance.*



Please note: the two letters identify the two alleles of the SNP. For example G, C indicates that on one of the chromosomes you would find a G (a G/C base pair) and on other chromosome you would find a C (a C/G base pair).

i. What is the **most likely** mode of inheritance *(autosomal dominant/ autosomal recessive/ X linked dominant/ X linked recessive)* of this disease?

X linked recessive

ii. Identify the SNP that is tightly linked with the disease allele.

G

Question 1 continued

iii. What is the most likely genotype at the **Z locus** of Individuals 1 and 2 in this pedigree? *Note: Use the symbol X^D, X^d, D or d where appropriate. In each case, use the letter "D" to represent the allele associated with the dominant phenotype and 'd" to represent the allele associated with the recessive phenotype.*

Individual 1: *X^dY* Individual 2: *X^DY*

iv. What is the SNP at the **Z locus** of Individuals 1 and 2 in this pedigree?

Individual 1: *G* Individual 2: *A*

Question 2

You plan to compare the sequence of the wild- type and the mutant alleles of Gene Z.

a) The following is the DNA sequence of the wild type allele of Gene Z that you want to amplify using the polymerase chain reaction (PCR).

```
5'CTCGAGGTGAATATGAAAG----|        |---CATTTGGCGCGTAATCGATA3'
                         | Gene Z |
3'GAGCTCCACTTATACTTTC----|        |---GTAAACCGCGCATTAGCTAT5'
```

i. Which of the following **sets of primers** would you use for PCR?

ii. Which of the following **sets of primers** would you use for PCR?

```
  Set1: 5'TCGGGGTGGATATGCA3'  and 3'AAGCGCGCAGTAGCTAT5'
  Set2: 5'TACACTTATACTTTC3'   and 3'GTAAACCGCGCATTAG5'
[✓ Set3:] 5'CTCGAGGTGAATAT3'  and  3'CCGCGCATTAGCTAT5'
  Set4: 5'GAGTTACACTTATAC3'   and 3'TGGCGAGTAATCGATA5'
[✓ Set5:] 5'CTCGAGGTGAATATGA3' and 3'GTAAACCGCGCATTAGC5'
```

iii. If you start with one molecule of DNA and allow the amplification cycle to be repeated 40 times, approximately how many molecules of DNA will you have at the end of 40 cycles? *Note: You may assume that you have sufficient primers and dNTPs for all the cycles.*

You would approximately have 2^{40} DNA molecules.

iv. You use the Taq DNA polymerase enzyme in PCR. What advantage does this enzyme have over the eukaryotic DNA polymerase?

Taq DNA polymerase (originally derived from Thermus aquaticus bacteria of hot springs) unlike the regular DNA polymerase, is highly thermo stable and remains functional even at a high temperature condition i.e. 95^0C required during each PCR cycle.

b) You now decide to use dideoxy-sequencing method to sequence the PCR product from both the normal and affected individuals. Briefly outline the basic principle used in dideoxy sequencing.

This method uses the dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. The ddNTPs lack a 3'- OH group required for the formation of a phosphodiester bond between two nucleotides during DNA strand elongation. The method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, radiolabeled deoxynucleotides (dNTPs) and ddNTPs that are added at lower concentration than the standard dNTPs to allow strand elongation sufficient for sequence analysis. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard dNTPs (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four ddNTPs (ddATP, ddGTP, ddCTP, or ddTTP). Incorporation of a ddNTPs into the nascent (elongating) DNA strand terminates DNA strand elongation resulting in DNA fragments of varying lengths that can then be resolved through DNA gel electrophoresis and the sequence can be read on an X –ray film. The terminal nucleotide base can be identified according to which ddNTP was added in the reaction giving that band. The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.

Question 2 continued

c) Using dideoxy sequencing, you derive the following sequence that corresponds to the **coding region** for amino acids 20-25 of the protein encoded by Gene Z.

5′ TTACCTAGCGTATGAAATC 3′

i. Write the DNA sequence for the region that corresponds to the amino acids 20-25 of the protein encoded by Gene Z and label the 5′ and 3′ ends of both strands.

5′… TTACCTAGCGTATGAAATC …3′

3′… AATGGATCGCATACTTTAG …5′

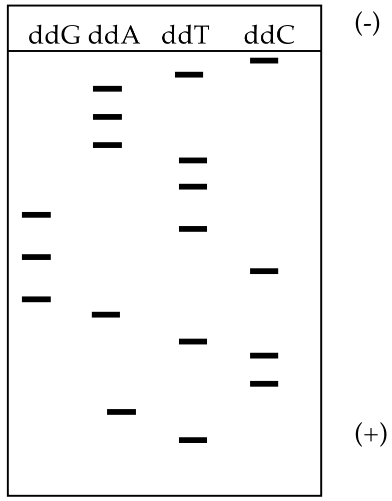
ii. Write the mRNA sequence for the region that corresponds to the amino acids 20-25 of the protein encoded by Gene Z and label its 5′ and 3′ ends.

5′… UUACCUAGCGUAUGAAAUC …3′

iii. Write the sequence of the amino acids 20-25 of the protein encoded by Gene Z and label its N and C termini.

N…tyr^{20}-leu^{21}-ala^{22}-tyr^{23}-glu^{24}-ile^{25}…C

d) You then sequence the mutant allele of Gene Z from an affected individual and observe the following pattern for the **coding region** for amino acids 20-25 of the protein.



Based on the DNA sequence that you derived and assuming that this sequence represents the coding strand, complete the following table.

Name and position (i.e. gly^{20}) of the amino acid…		Type of point mutation (nonsense/missense/ frameshift/silent)?
in the wild-type version of the protein encoded by Gene Z	in the mutant version of the protein encoded by the mutant allele of Gene Z	
Tyr^{23}	*cys^{23}*	*Missense*
glu^{24}	*The codon for glu^{24} is now changed to a stop codon*	*Nonsense*

Question 3

Using yeast as the system, you want to study the genes responsible for arginine biosynthesis. You obtain several arginine auxotroph; arg1-, arg2-. arg3-, arg4- and arg5-. Each of these auxotrophs is defective in only a single enzyme in the arginine synthesis pathway.

The table below indicates the results of crossing haploid cells of the indicated genotype and testing whether the resulting diploid cell can grow in the absence of arginine (denoted as +) or requires arginine (denoted as -).

	arg1-	arg2-	arg 3-	arg4-	arg5-
arg5-	+	-	-	+	-

Question 3 continued

a) You want to clone the arg5 gene using cloning by complementation technique. You begin by constructing a **yeast genomic library in *E. coli***. Which yeast auxotroph(s) (*arg1- , ar2-, arg3-, arg4- and arg5-*) could you choose as the **donor** for the genomic DNA? **Explain**.

You would either use arg1- or arg4- mutants since they have the wild-type version of the arg5 gene, which can complement the mutation in arg5-mutant (as shown by the + signs in the table above).

b) You successfully prepare **yeast genomic DNA** and need to choose a vector that will allow you to…

1) create the yeast genomic library in *E. coli,*
2) use the library to transform arg5 yeast auxotrophs and
3) clone by complementation the gene that can restore the yeast arg5 prototrophy

You need to choose a vector that will allow you to complete this experiment. From the options of the plasmids below, circle the plasmid that has the **<u>minimum features</u>** that are required to execute the plan outlined above.

[✓ i.] Bacterial ori, bacterial selection marker, yeast ori, restriction enzyme site.
ii. Bacterial ori, bacterial selection marker, yeast ori, bacterial promoter, restriction enzyme site
iii. Bacterial ori, bacterial promoter, yeast promoter, bacterial selection marker, restriction enzyme site
iv. Bacterial ori, bacterial selection marker, yeast ori, yeast selection marker, restriction enzyme site
v. Bacterial ori, bacterial promoter, yeast promoter, bacterial selection marker, restriction enzyme site
vi. Bacterial ori, bacterial selection marker, restriction enzyme site

c) You digest both the yeast genomic DNA and many copies of the vector with the EcoR1 restriction enzyme. You mix the genomic fragments with the cut vectors and add DNA ligase. You then transform *E. coli* cells with the ligation mix.

i. During your transformation, you hope to have a ratio of bacterial cells to recombinant plasmid of about 10-cells/recombinant vector. Why would you want more bacterial cells than recombinant vectors?

All the bacterial cells will not necessarily take up a recombinant vector. By having many more bacterial cells, you increase the likelihood that all the recombinant vectors are taken up by the bacterial cells. Alternatively, this strategy also helps to minimize the number of bacteria that may receive multiple plasmids.

ii. What growth medium would you use to distinguish the bacterial colonies that did not carry a recombinant vector from the ones that did carry a recombinant vector?

Your growth medium will depend on the selection marker gene that is a part of the plasmid that you use as a vector to clone the arg5 gene. If the plasmid has an ampicillin resistant gene (amp^r) you will plate the bacterial cells on medium containing ampicillin. Only those bacterial cells that receive the plasmid will survive and form colonies in this growth medium.

d) You successfully create a yeast genomic library in *E. coli* cells. Outline how you would use this library to clone by complementation the gene that can restore the arg5 prototrophy in arg5 auxotrophs.

Isolate recombinant plasmids, containing different yeast genomic fragments, from the bacterial colonies. Transform the arg- yeast auxotrophs with the recombinant plasmids. Plate the transformed yeasts into minimal media. Yeasts that grow on minimal media will carry a gene that functions in the arginine biosynthetic pathway and can restore the arg prototropy.

e) You transform the arg5- yeast auxotrophs with the recombinant vector that has the wild-type arg5 gene. What medium could you use in petri plate to distinguish between the arg5- auxotrophs from the arg5 prototrophs?

You will use the medium that lacks arginine.

f) You successfully identify a recombinant vector that carries the wild-type arg5 gene. Assuming that the yeast and bacterial versions of the protein encoded by the arg5 gene are identical, can you use this recombinant vector to rescue a bacterial cell that is also an arg5 auxotroph (Yes/No)? **Explain**.

No, you will not be able to rescue the bacterial cell that is also an arg5 auxotroph because the putative arg5 gene cannot be spliced in bacteria. Or the vector used to make the genomic DNA library lacks a bacterial promoter, and thus the protein will not be expressed in bacteria.

Question 4

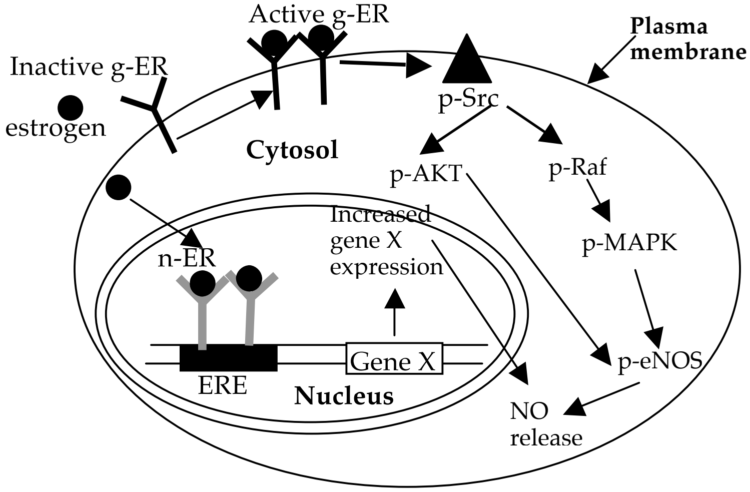
The steroid hormone estrogen, in pre-menopausal women, puts them at a lower risk of developing cardiovascular disorders compared to the age matched males.

Estrogen mediates its cardio-protective effects by binding to G-protein coupled receptors (g-ER) on cell surface that activate intracellular signaling cascades and result in the activation of endothelial nitric oxide synthetase enzyme (eNOS) through phosphorylation. The active eNOS is involved in production of nitric oxide (NO) that acts as a blood vessel dilator to mediate its cardio- protective effects.

Estrogen can also act by binding to specific nuclear hormone receptors (n- ER). These n–ER, on binding to estrogen, can act as co-transcription factors that bind estrogen response element (ERE) on the DNA and enhance the expression of Gene X that also enhances the NO production.

For simplicity you may assume that the ***activation of either of these pathways (g-ER or n-ER mediated) is sufficient for cardiac protection****.*

A simplified schematic of the signaling pathways activated by estrogen is shown below:



Note:

- *Estrogen response element is abbreviated as ERE.*
- *All other proteins, except eNOS, are* ***kinases****.*
- *All proteins shown in the cytosol are activated through phosphorylation.*
- *The membrane estrogen receptors (g-ER) are shown as black and the nuclear hormone receptors (n-ER) that also bind to estrogen are shown as gray.*

a) Complete the following table for each of the following perturbations or mutations.

Perturbations/mutations	**Cardio-protection by estrogen (Yes/ No)?**	**Receptor (g-ER/n-ER) responsible for cardio- protective effects of estrogen?**	**Explain.**
Src kinase is never phosphorylated.	*Yes*	*n-ER*	*Src kinase must be phosphorylated for the activation of e-NOS through the Src-Raf-MAPK or Src-pAKT pathway. So this mutation will prevent the activation of e-NOS through these pathways. However, estrogen will still be able to mediate its cardio- protective effects by binding to n-ER and enhancing the Gene X expression that will promote NO production that has cardio- protective effects.*
Mutation that prevents ERE from binding to n-ER.	*Yes*	*g-ER*	*This mutation will prevent the gene X mediated NO production. However, the g-ER mediated pathway will still operate through which estrogen will mediate its cardio- protective effects.*
Cells are treated with a membrane impermeable form of estrogen.	*Yes*	*g-ER*	*This mutation will prevent the g-ER mediated activation of eNOS. However, the n-ER mediated pathway will still operate through which estrogen will mediate its cardio- protective effects.*

b) The schematic shows that estrogen from the surrounding environment can directly enter the cell's cytoplasm and nucleus. Explain why is this possible.

Estrogen is a steroid hormone that is synthesized using cholesterol as a precursor. This hormone, similar to cholesterol is highly hydrophobic and hence can easily diffuse through the hydrophobic environment provided by the lipid bilayer of the membrane.

Question 5

Lymphomas are broadly classified as diffuse large B cell lymphomas (DLBCL) and follicular lymphomas (FL). Prognostic models based on pre-treatment characteristics are currently used to predict the treatment outcome. However, these models neither identify the molecular basis of clinical heterogeneity nor specific therapeutic targets.

a) One strategy is to use microarray analysis to understand the gene expression pattern of the disease. Briefly describe how microarrays may be used to understand the expression pattern of genes.
A microarray consists of an array series of thousands of spots of specific DNA sequences known as probes. These can be short section of a gene or other DNA elements that specifically hybridize/undergo complementary base pairing with specific DNA or RNA sequences under high stringency conditions. This hybridization is then detected and quantified to determine the relative abundance of nucleic acid sequences in the target.

b) Cluster the patients into two distinct classes; Class I and Class II. **Explain** how you made this classification.
Based on the information provided patients 1-42 belong to Class 1 and 43-77 belong to Class II based on the highest branching in the clustering tree.

c) A patient comes in with tumor. After taking samples of patient's tumor and running the appropriate diagnostic tests you find that the tumor showed a high expression of Bcl-2 (*feature U29680 at*), Cyclin B1 (*feature M25753 at*) and lactate dehydrogenase (LDH, *feature X02152 at*), which are the known protein markers for DLBCL. Which of the two classes (class I or class II) that you identified in part (b) represents the DLBCL patients?
The Class I patients have DLBCL since they show a higher expression of Bcl-2 and LDH compared to the patients that belong to class II.

d) Cathepsin genes (*feature M63138 at*) encode proteins that promote invasion by the cancer cells and metastasis of cancer. Based on this observation which of the two classes of lymphomas is more aggressive?
Since cathepsin gene shows a higher expression level in DLBCL (Class 1) patients, this seems to be a more aggressive form of lymphoma compared to FL.

e) Name the two genes that may be regarded as the signature sequence/markers for DLBL.
XO3689_at (mRNA fragment of elongation factor TU (N terminus), score =6.892) and Z21966_at (POU6F1 POU homeobox protein, score = 6.294).

f) Recent studies indicate that the gene encodes PDE4B gene product (*feature L20971_at*) that promotes cellular proliferation and is therefore associated with a poor prognosis in DLBCL patients. Based solely on this information, do the DLBCL patients included in this study most likely have poor or good prognosis? **Explain** your choice.
The microarray data show that the score for L20971_at gene expression is -1.238, which reflects its low expression level. So based solely on this information, the DLBCL patients in this study appear to have a good prognosis.