BIOL 426/BIOL626

Final

December 13, 2018

8:00 AM to 9:50 AM

Be sure to put your name at the top of each page!

The examination will consist of five questions:

Answer all parts of Question 1	15 pts
Answer any three of Questions 2–5	60 pts
Answer either of Question 6 or 7	25 pts
Total:	100 pts

Please be sure to answer the correct number of questions. Remember that partial credit will be given generously, ***so don't leave questions blank!***

Also, please try to be *concise* (there will be no extra credit for long–winded answers) and use the available space to answer the questions.

Take into account the number of points allotted to each answer when budgeting your time!

In multi-part questions remember to clearly identify each part of your answer

Before beginning to answer the questions, please enter your name on EACH page of the exam

1. (15 pts) **Short Identifications.** In the space provided define the following terms in a few sentences. Provide enough information to *uniquely identify* the term!

a. (3 pts) RFLP

"Restriction fragment length polymorphism" a difference in the size of restriction fragment(s) that is caused by a mutation that is prevalent in a population. The variation in size of the fragment can be used as a visible genetic marker to identify alternative alleles of genes of interest.

b. (3 pts) Transcriptome vs proteome

The transcriptome is the collection of all RNA molecules in a cell. The proteome is the sum of all proteins present in a cell.

c. (3 pts) CpG island

CpG islands are regions in chromosomes that have a higher than average density of the sequence 5'-CG-3', a palindromic sequence that can be subject to methylation to 5-methylcytosine. The islands tend to encompass promoter regions and the CpG sequences in them are usually less densely methylated than those in regions where CpG is less dense.

d. (3 pts) Totipotency

Totipotency is a developmental state in which a cell (a stem cell) is able to adopt any of the differentiated cells normally found in the organism.

e. (3 pts) CRISPR

"Clustered Regularly Interspaced Short Palindromic Repeats". The CRISPR sequences are a kind of adaptive immune system for certain bacterial species. The DNA sequences encode guide RNAs that along with enzymes like Cas9 nuclease target foreign DNAs for double strand cutting. The CRISPR-Cas9 (and some other) systems have been adapted to allow directed cutting of target DNAs using complementarity to a single guide RNA (sgRNA). More advanced versions fuse inactive Cas9 nuclease to other proteins such as transcriptional activation or repression proteins to modulate transcription of adjacent genes.

Answer ***any three of the next four*** questions (#2–5)

2. (20 pts)

   The paper by Glazov et al. described the use of whole exome sequencing to identify a novel disease causing gene for anauxetic displasia, *POP1*.

   a. (10 pts) Why is whole exome sequencing the method of choice for identifying such novel disease causing genes. In particular, what advantage does it have over other methods including using SNP arrays or next generation whole genome sequencing.

   b. (10 pts) How did Glazov et al. use evidence of evolutionary conservation in the family of genes to which the novel disease gene belongs to focus on *POP1* for more detailed analysis? How did the functional relationship between the other known gene, *FMRP*, help solidify the identification of *POP1* as a causative gene for the disease?

   a. Whole exome sequencing allows a fast and relatively inexpensive search for sequence differences in the genome of cells carrying disease causing mutations. Sequence differences found in this way can be further tested to determine if they cause the disease. The major disadvantage of SNP arrays is that the causative mutations are not frequent enough in the population to be in the array, so SNP screening cannot find them. Whole genome next generation sequencing would actually find genes not available for whole exome sequencing because they fall outside coding regions, but it is too expensive and too slow to be used for this purpose.

   b. Glazov use a "SIFT" algorithm (you didn't need to remember that name) to look for changes to evolutionarily conserved nucleotides on the theory that mutations capable of causing a significant reduction in gene function should alter such conserved nucleotides. The POP1 mutation was such a change. Pop1 protein was known to form a molecular complex with the product of the RMRP gene (sorry for the typo in the question!) encodes the RNA component of RNAse MRP and Pop1 is a protein component of the same complex. This complex has an essential function in post-transcriptional processing of ribosomal RNAs.

3. (20 pts)

The Smith et al. paper examined the changes in DNA methylation from oocytes and sperm, through formation of the zygote, early development and finally the pre-implantation embryo.

a. (10 pts) Describe the general state of methylation of the oocyte and the sperm. How does methylation of the maternal and paternal chromosomes change in the zygote and then through early development up to the pre-implantation embryo.

b. (10 pts) What nucleoside is methylated in this system? What was the method used by Smith et al. to map those methylated nucleosides in the genome? How does that method distinguish between the methylated and unmethylated form of that nucleoside?

a. Oocyte DNA is globally hypomethylated compared with sperm DNA. In the zygote the paternal chromosomes (from the sperm) are strongly demethylated and the maternal chromosomes less strongly demethylated. The zygotic DNA overall is hypomethylated similar to the oocyte. The methylation of both types of chromosomes increases during early development until the pre-implantation embryo when the methylation resembles that of adult somatic cells.

b. The methylation is of cytosine to 5-methyl cytosine. These methylated nucelosides can be identified by bisulfite sequencing. When treated with bisulfite, cytosine is converted to uracil but 5-methylcytosine is much less sensitive so they remain essentially unchanged. Sequencing through a region that includes methylated bases, only the base pairs that involve 5-methylcytosines appear as C-G pairs in the sequencing reaction and those involve cytosines appear as U-A pairs. So, any C-G pairs found are by definition the sites where 5-methyl cytosine had been present.

4. (20 pts)
5. You would like to use CRISPR-Cas9 to test the role of a gene of interest in a phenotype you're interested in studying.
   a. (10 pts) In designing the single guide RNA (sgRNA) what sequence features would you need to have in it to be able to target your gene?
   b. (10 pts) You would like to reduce or eliminate the expression of your gene. Describe two mechanistically different ways you could achieve that end using CRISPR-Cas9 technology. (Assume that the gene you wish to study is not essential for life.)

a. The portion of the sgRNA that has to be designed is the region that is complementary to the target sequence in the genome. That target sequence must begin at the 3' end with the protospacer adjacent motif (PAM), which for Cas9 is 5'-NGG-3'. This sequence must be present in the genome at the target site. Adjacent to the PAM is the "seed sequence", which must be entirely complementary to the target (the rest of the sgRNA to the 5' end can be slightly non-complementary..

b. Here's what I was thinking of. You can design a system using an sgRNA and active Cas9 nuclease and it will cut the target DNA and after non-homologous end joining (or introduction of a mutant sequence using a DNA carrying that mutation and homology directed repair) the gene is mutated to lose its activity. Alternatively, you could use the same sgRNA and a catalytically inactive Cas9 fused to a transcriptional repression protein that either directly blocks transcription or modifies the chromatin in the region to reduce transcription (either would have been fine). In this case, the gene is not mutated but it isn't transcribed.

Other acceptable answers could have described alternative methods of delivery of the sgRNA-Cas9 complex, for example using a retroviral vector system, or any modification of the standard sgRNA-Cas9 system described above.

5. (20 pts)
   a. (10 pts) Describe the process of genome wide association studies (GWAS) to map particular genetic phenotypes fo regions of the genome. How are single nucleotide polymorphisms (SNPs) involved in that analysis?
   b. (10 pts) Describe how SNPs are used to map regions of the genome that have undergone loss of heterozygosity?

*a. Genome wide association studies look for either of two common allele sequences in the population at all SNP sites across a genome. These are called the A and B alleles. The assay uses hybridization to probes that carry either of each common SNP types so it can rapidly identify the presence of many SNPs across the whole genome or any genomic region.**

b. In regions where there has been loss of heterozygosity (LOH), the whole region can carry only one of the two alleles (either A or B) and in this region only the DNA complementary to the single allele is present. This contrast to the homozygous diploid regions that can have one, the other or both SNPs at each position. By simply looking at the plot of alleles present across a region one can see the presence of LOH by the complete absence of the heterozygous condition (both A and B are present).

* I notice that I didn't specify that GWAS is done across a large group of individuals some of whom have the genetic disease being tested and some who don't. The SNPs that are more prevalent in the disease patients than in the controls are the ones that indicate a gene or genes that are associated with the disease.

Answer ***either one of the next two*** questions (#6-7)

6. (25 pts)

   You are interested in developing a mouse model system to understand the role of a protein kinase, Kin1, in the progression of pancreatic cancer. You have data showing that the expression of Kin1 protein is increased in diseased tissue from its level in normal pancreas so you think that the uncontrolled expression of Kin1 may contribute to development of the cancer.

   a. (10 pts) Describe the general approach you would take to construct a transgenic mouse model that would allow you to study whether overexpression of Kin1 is sufficient to cause cancer. What controls would you want to do to show that Kin1 kinase activity is necessary to cause cancer?

   b. (10 pts) Exploiting the system you describe in part (a) above, you would like to test if a small set of candidate proteins are targeted by Kin1 for phosphorylation. Describe an experiment to test this question (assuming you have any specific tools you might need).

   c. (5 pts) You are interested if another set of kinases—Kin2, Kin3 or Kin4—are part of a multi-kinase cascade between Kin1 and the target proteins. This putative cascade would involve each kinase being phosphorylated by an upstream kinase, activating its phosphorylation of a downstream kinase, leading eventually to phosphorylation of the target protein. Describe an experiment to determine if any or all of Kin2, Kin3 or Kin4 are members of this hypothetical cascade.

   a. The approach is to build a construct in which Kin1 is expressed only in the prostate and only when you want it to be. That could be done in several ways (and I will accept any that makes sense) but one way is to express the Tet-on system under the control of a prostate-specific promoter and the Kin1 gene under the control of the Tet-on activator protein. You then create a transgenic mouse carrying these constructs and induce high expression of Kin1 by addition of tetracycline. Controls would include introducing only the prostate-specific Tet-on, or only the Kin1 under control of Tet-on; neither will lead to overexpression. Another control would be to introduce a gene encoding a catalytically inactive Kin1 protein so though it would be overproduced it couldn't have any further effect. Only the combination of the two wild type proteins would result in increased cancer incidence if Kin1 does contribute to cancer progression.

   b. Assuming that you have antibodies specific to the several putative downstream targets of Kin1, you would induce Kin1 overexpression and monitor the target proteins for evidence of phosphorylation. This could be done by Western blotting 2D gels run to show phosphorylation by its shifting the pK of the protein or you could purify the targets by immunoprecipitation and look for the presence of phosphorylation by a radioactive tracer (^{32}P).

   c. To show that Kin2, Kin3 or Kin4 are intermediates in the cascade resulting in increased cancer progression, you would mutagenize these genes (for example, using CRISPR) and test if the mutations eliminate increased cancer progression when Kin1 is overproduced. Any second kinase that is between Kin1 and whatever protein(s) are targeted to induce cancer should eliminate the increased cancer progression. Those that are not will have no effect.

7. (25 pts)
   a. (10 pts) Describe <u>three methods</u> for studying transcription and its control across the genome, at least one of which can be used to study all or nearly all coding RNAs expressed in the cell? Rank the three methods in terms of their coverage of those RNAs, the quantitative nature of the method and the ability to study both high and low abundance RNAs.
   b. (10 pts) If you were interested in studying the regulation of transcription in a previously undescribed bacterium, what method would be simplest for you use to do that? What other study or studies would help you to analyze the data from that transcriptional study?
   c. (5 pts) Which of these methods could you most easily use to test for evidence of alternative RNA splicing across the genome? Describe how that type of analysis would be done and what evidence the experiment would produce to show that alternative splicing occurred.

a. Among the methods you could have described are subtractive hybridization, differential screening, differential display, expressed sequence tag (EST) arrays, and RNAseq. These methods are all described in the Transcriptomics class presentation. Basically, I'm looking for sufficient information to distinguish each of these from the others. These techniques essentially in the order mentioned increasingly provide coverage and quantitative measures. RNAseq is the most effective for both and is the only one of the approaches that really can give good results on both high and low abundance RNAs.

b. All of the approaches require a significant amount of work to produce tools capable of interrogating a previously undescribed bacterium. The only method that would easily allow you to do this would be RNAseq since by simply purifying RNA from the organism you can get complete transcriptome coverage. To make this work most effectively, a sequence of the bacterial genome could be completed by next generation sequencing so that you would be able to identify all the genes present and compare them with those identified by RNAseq.

c. Once again, only RNAseq is able to give you information about RNA splicing without having detailed knowledge of alternative spliced forms. By doing RNAseq and comparing the sequences obtained to the genome sequence a splice joint is a sequence at which two non-adjacent genomic sequences are brought together in a single RNA sequence. In addition, at the junction of these two sequences you would find sequence motifs typical at the sites of splicing.