

Genomics Midterm Thursday, March 25, 2010

NAME: _____ STUDENT ID: _____

Take-home exam: Due Thursday, April 1, 2010.

Return an editable WORD document with answers by email to David Pollock (David.Pollock@ucdenver.edu) by **5:00 pm on April 1, 2010**. Please use “**Genomics Midterm**” as the subject.

I will send out a **blind-copied confirmation** to everyone from whom I receive a completed exam. If you do not receive this confirmation, please call to find out why not.

All answers to questions should be typed and submitted as a WORD document. Include your name in the name of the document, e.g., **DPollock_GenomicsMidterm.doc**

Answers to each question should be **concise yet informative and ~1/2 page**. For each question you answer, you should go over your notes and any reading related to the relevant lecture(s) thoroughly so that you fully understand the material and this understanding is reflected in your answer. Note, the purpose of this midterm is to get you to go back over your lecture notes and review key points that you should take away from the class. We’re also hoping to encourage you to think about and digest the information a bit, and demonstrate that you have done so.

Please **answer only 8 out of 10** of the following questions (your choice):

1. Why was “physical” mapping of genomes necessary for the clone-based DNA sequencing approach that has provided nearly all genome sequences? Will physical mapping of genomes remain necessary in this age of high-throughput whole-genome shotgun sequencing? Why or why not?
2. When sequencing diploid individuals, each base must have 8x or greater coverage to have confidence in your base calling. If you sequence a genome at an average of 20x coverage, what fraction of the bases are at 8x or greater coverage? Use the Poisson distribution: $f(n; \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$, to calculate your answer. (n = # of events = # of times a given base is sequenced; λ = mean # of events = average sequence coverage).
3. Describe the concept of “exaptation” with regard to transposable elements. Pick a class of transposable elements (e.g., DNA transposons, retrotransposons) and discuss what features make this class of transposable elements particularly well-suited to be exapted to form a new regulatory network?

4. Now that we have over 1000 complete human genome sequences, what is the utility of sequencing more genomes from different species, some closely related to humans, others distantly related. Specifically, provide three brief examples of how comparative genomics (and increased knowledge of genomic diversity) contributes to a greater understanding of the human genome and human biology. For each example, also identify which types of species (how closely related to humans, in a very rough sense) are most important for each comparison discussed.

5. Segmental Duplications (SDs) are a class of repetitive DNA in the human genome. How are SDs classified? Briefly explain how SDs may have lead to the creation of new genes.

6. Using the NCBI and other databases, please write a summary of all that is known about the following human sequence. If available, include at least information about taxonomic distribution of homologs, genomic location, protein structure, expression, functions (including pathways and phenotypes) and relevant variants.

```
TAGAGATCAG GCCTCGGGAG ACCCCCACCC TGTGCTCCCC ATGTC CCTTG CCTGCACCAT
GAAGTTGAGG GAGGGAGAAG GCCTGGCTCT CCCGGAGTCA GGCAGGCGGC AGTGGGACCC
AGAGCCCAGG ATGCTGCCAG GCCAAGCAGC AGCAACACTC ACCGCCGTGT GGCTGGCTTT
GCCGGTGAGG ATGCTGCGGG CCTTCTTTT GGTCATGTTG AAGTTTTTCA GGTAGGCATT
GTAGATGTGC TTGGAGAAGG CCTTCAGGTC GGCCACCTGT GGGTTGTA CT GGCTCCCCCTC
TTTGCA GTCA GCCCTGCCAC CAGCTTCCTC TTCTCAGCCT CCGGCATCCG ACCAAAACGG
ATAGCTGCAC AGGGAAGGGG GCAGTCAGCA AGGAGCCCAG GCAGGCCCCA GCACCTCTGA
CATCCCCATC CCTTACAGG TGCA TGCGCC AAATCCTTTT GGAGCCTAAG GCCCCAGAA
GCTCTAGAGT CAGCAAGGTA GGAATGATGG GGTGGCCCCT CCAGATCGCA AGCTCCACAA
GATGTGATTT TTTTTTTTTT TTGGTGTATC TTTGCAAGAG CAGTGATTTT GTCTGTTTTG
```

7. Based on NCBI build 36/hg18 locate the region Chr18:10,000,000-13,000,000 and then answer the following questions:
 - a) What types of segmental duplications (SDs) can be found within this region?

 - b) Identify the SDs most likely to mediate non-allelic homologous recombination (NAHR). What are their co-ordinates, size and percent sequence identity?

 - c) What is the approximate size of the deletion/duplication resulting from NAHR mediated by these SDs? How many genes are involved?

8. First, think about (and review) what you have learned about how transcription factors interact with each other in animal genomes (e.g., flies or humans) and bind to hundreds of different binding sites across these genomes. Then speculate (in the context of the lecture material) how the entire set of transcription factor binding sites might evolve depending upon the sequence specificity of a transcription factor, and also how this repertoire of transcription factor binding sites might “react” over time, to a transcription factor amino acid replacement that affects binding specificity.

9. A UC Berkeley professor was quoted as saying (a long time ago) that “RNA” stands for “Really Not Applicable”. Refute this scurrilous statement based on what you have learned about RNA expression analysis and “RNA Genomics” from the course lectures. Discuss the advantages and disadvantages of at least two different approaches to gene expression analysis (e.g, EST, microarrays, RNAseq).

10. Although RNA may actually now be seen to be rather interesting, discuss (based on the proteomics lecture) why it may be important to analyze proteins as well. Also, what are two important modern techniques for protein quantitation (please list a pro and con for each)?