Bi188 Final Examination
Spring 2011

Due **Sunday, June 5th at noon** for seniors and **Saturday, June 11th at noon** for everyone else (electronic copy emailed to kfisher@caltech.edu, georgi@caltech.edu, and woldb@caltech.edu).

The exam is **closed-book**, **closed-notes**, and **closed-internet**. You have 3 hours to complete the exam. *If you draw out an answer and take time getting it back into electronic form (i.e. scanning/drawing in Microsoft Paint/etc.), you do not have to account for this in the 3 hours.*

Answer all of the first 5 questions. There is an optional sixth bonus question at the end for extra credit.

Reminder: the exam is **CLOSED-BOOK**. Read no further until you are ready to take it!

Question 1

Complex genetic diseases such as type2 diabetes remain a major challenge to unravel.

A. How would you measure heritability in a given population, assuming you could get whatever samples/families you wanted? Do you expect that this measurement, performed now by you and performed 100 years ago would give the same result, and if not, why not? How would you expect the heritability value h2 to have changed?

B. Genome-wide association studies are now plausible using dense million-SNP arrays and access to large study populations (10,000 to 100,000 individuals) consented for GWAS. Given better and better assay methods, why would you NOT want to expand the assay to include all discovered SNPs?

C. There are now ~17 GWAS putative positives for type2 diabetes from studies conducted around the world. What do you expect the allele frequencies (minimally) to be in the population where they were discovered? Assume that best practices were used in all studies and that a substantial followup was conducted in an independent sample from the same population. How do you expect the association significance to change in the second study? Why?

D. You are drawn to a new project in which there is much excitement because you have a promising novel resource. The resource is a set of three unrelated families in which an extreme and very similar version of type2 diabetes is observed. It is extreme because it manifests strongly at an early age, but also depends strongly on body weight (high BMI is a phenotype) and clarifies that this is not classic type1 childhood diabetes. In one family, you have three affected siblings and two unaffected ones. In the other two families you have two affecteds and one affected - respectively.

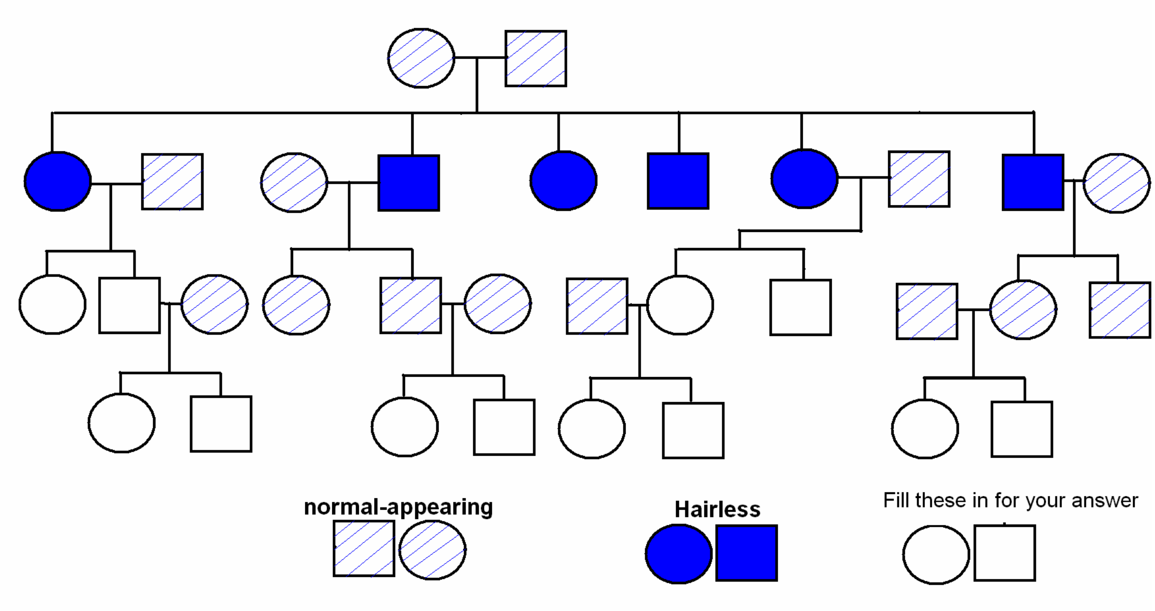
   1. How would you use DNA samples from these families to identify the gene?
   2. What would you do to confirm that your identified gene strongly predisposes to diabetes and is not merely a statistical outlier?
   3. Do you expect the alleles you find to be the same or different from each other and why?
   4. Do you expect future GWAS studies to find it? Why?

Question 2

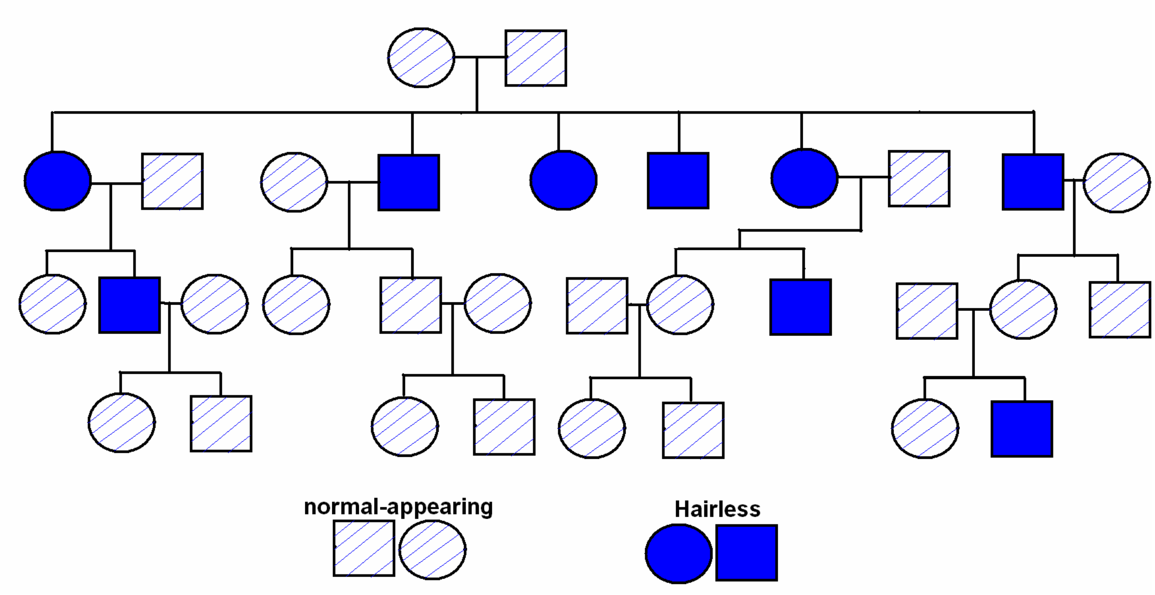
Years ago, a number of siblings (three boys and three girls) were born completely hairless. The parents were both normal appearing. The initial thought was that there was some toxin in the well at the family ranch. Some hairless grandchildren (who are of both sexes) were born to the hairless girls who had moved away to far off cities (and also to one daughter who stayed at the ranch). No children of the hairless men in the family - either men who stayed near the ranch compound or moved away - had hairless children.

a. Why would you suspect that genetic imprinting is involved? What is the most likely explanation for it not affecting either of the founder parents?

b. Assuming the cause of hairlessness is genetic imprinting, complete the following kindred tree showing the founder woman who is herself unaffected and down to her great-grandchildren's generation:



c. What would you suspect was the cause if the hairless women's sons were hairless but their daughters were normal appearing (as in the below example)? Note that one of the hairless man's grandchildren is also affected.

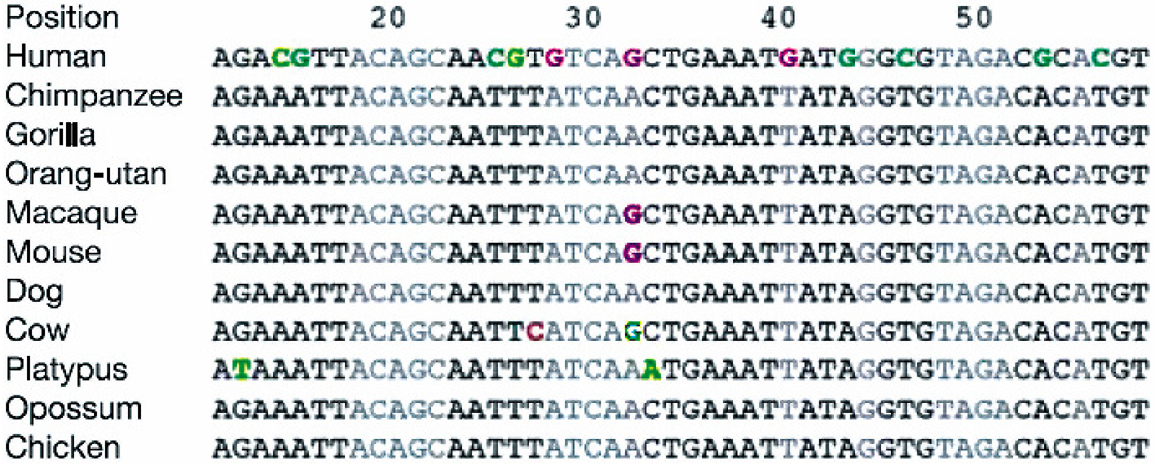


d. How would you determine both locally and globally whether DNA methylation were involved in this disorder? You would have access to any tissue or DNA samples you needed. This part should be doable in 3-5 sentences.

e. Using a functional assay of your choice (you need to specify what each experiment would tell you and what alternate possibility it would eliminate) state how you would determine if imprinting is the cause. This part should be doable in 3-5 sentences.

Question 3

In 2006, David Haussler and colleagues at the University of California, Santa Cruz published a paper (for your future interest beyond the exam, it is at http://www.ncbi.nlm.nih.gov/pubmed/16915236) describing their efforts to identify regions of the human genome that show an unusually high rate of substitution since the divergence from our last common ancestor shared with chimpanzees and bonobos. Based on these analyses, they uncovered 49 such regions in the human genome. Below, we provide the sequence of the "human accelerated region" (HAR1) that shows that the most dramatic difference relative to other non-human primates and other mammals. Green and purple-colored bases differ from the consensus sequence. It was subsequently determined that HAR1 is part of the multi-exon *HAR1F* gene that expresses a non-coding RNA in human and non-human primate cells.



```
Position                 20        30        40        50
Human       AGACGTTACAGCAACGTGTCAGCTGAAATGATGGGCGTAGACGCACGT
Chimpanzee  AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Gorilla     AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Orang-utan  AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Macaque     AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
Mouse       AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
Dog         AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Cow         AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT
Platypus    ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT
Opossum     AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Chicken     AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
```

Based on the information provided in the primate genomics lecture, please say what data and analysis from follow-up study material below could be used to address hypotheses concerning (i) the timing of the emergence of the accelerated nucleotide evolution of the HAR1 region in the human lineage and/or (ii) the potential functional role(s) the HAR1 region might have played in human evolution. In your answers, please highlight the major strength and weakness of each approach.

Do the first one and select two others. Each part should be doable in 3-5 sentences.

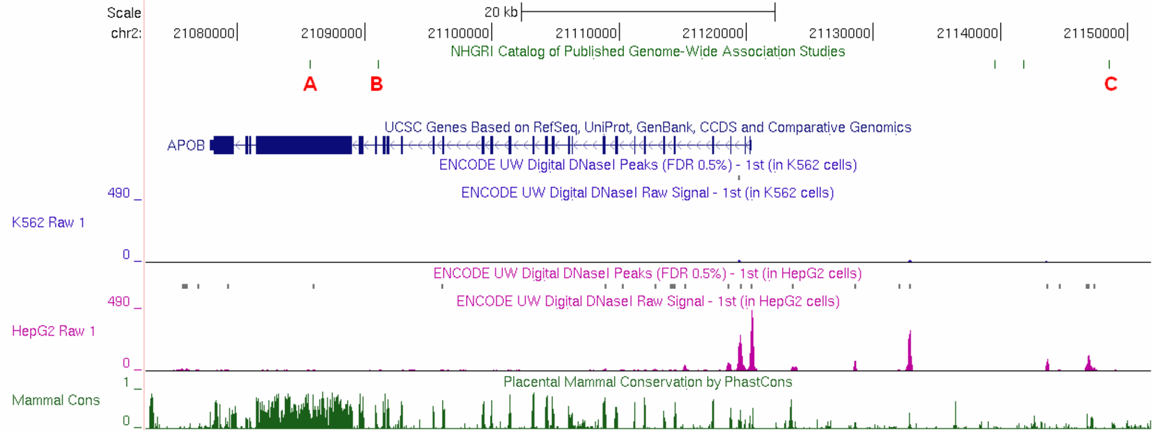
1. Information about the frequency of single nucleotide polymorphisms in the region surrounding the *HAR1F* gene in 1000 humans from diverse geographic regions.
2. The sequence of the Neandertal *HAR1F* gene.
3. Transgenic mice engineered to express the *HAR1F* gene from humans or other mammals. Assume that all endogenous copies of the mouse *Har1F* gene have been removed from their genomes.
4. Brain tissue from adult human and non-human primates.

Question 4

Last year, a genome-wide association study was performed on >100,000 individuals of European ancestry in order to associate SNP's that correlate with abnormally high or abnormally low blood lipid levels. Along with about 100 other SNP's, three SNP's appeared near the gene APOB. APOB (apolipoprotein B) is a subunit of low density lipoproteins and is expressed only in liver and gut cells. Below, the APOB locus and the surrounding genomic region are shown. The three SNP's are labeled A, B, and C in red. Note that SNP B falls inside an intron. The other tracks are DNase hypersensitivity data for K562* and HepG2* and the mammalian PhastCons conservation track**. You can assume that the DNase data are normalized for total read numbers and thus a relatively fair comparison between the cell types.

*the origins of the ENCODE cell types shown are:
***K562** = myelogenous leukemia (blood cancer) cell line*
***HepG2** = hepatocyte carcinoma (liver cancer) cell line*

***PhastCons track: 1 means that location is completely conserved among placental mammals. 0 means it is unconserved among placental mammals.*



A. Give two ways a SNP could affect the proper function of a gene if it is that gene's exon. Next, give two ways a SNP could affect the proper function of a gene if it is NOT in the exon of that gene.
B. Do you expect the causation of abnormal blood lipid levels to necessarily be at one of the three SNP's (A, B, or C)? What besides direct causation would explain these three SNP's being significantly associated with abnormal blood lipid levels?
C. You want to determine whether each of the three SNP's do, in fact, directly cause abnormal blood lipid levels. For each of the three SNP's, state the most likely way that it could cause this based on the information you were provided. What experiments would you do to test each of those data-informed hypotheses (give 1-2 sentences for each experiment)?
D. Give the most likely reason for the difference between the K562 and HepG2 DNase data near APOB.

Question 5

A. Rhabdomyosarcoma is a malignant tumor that affects striated muscle. In alveolar rhabdomyosarcoma, 70% of cases are explained by a translocation occurs between the *FKHR* (forkhead) transcription factor gene on chromosome 13 and the *PAX3* transcription factor gene on chromosome 2. You have a modern-day RNA-Seq (100 million reads, each read being a mate-pair with 100bp on either end of a 300bp fragment) sample from a young patient's rhabdomyosarcoma tumor and an RNA-Seq sample from his normal muscle obtained at biopsy. What would you look for in the RNA-Seq data to see if the FKHR/PAX3 translocation had occurred in your young patient? Please clarify your answer by drawing a simplified map of the expected read distribution if the translocation occurred in an exon of FKHR and an exon of PAX3. What would this look like if the translocation occurred at an intron of FKHR and an intron of PAX3? We want you to be explicit about how you would map the sequence reads to the known genome, to known gene models, and any additional models you would provide.

B. You are researching hematopoiesis and you are fortunate enough to have gathered some white blood cell samples from a family consisting of two parents and one child. You make RNA-Seq samples of each of the three samples. Based on genotyping, you know that the father's alleles are ata/aca at c-Myc; the mother's ata/aga; and the child's aca/aga. The alleles differ by one particular SNP in the coding region of c-Myc so you can determine from the RNA-Seq data which allele produced which transcript. RPKM values for the transcription factor gene c-Myc are as follows:

	overall	**allele ata**	**allele aca**	**allele aga**
mother	191.05	38.21		152.84
father	83.47	43.74	39.73	
child	208.85		41.77	167.08

What are two substantially different mechanisms that would explain this asymmetric RNA expression? Might the mother and child's phenotype have any bearing on their susceptibility to leukemia? Explain.

Bonus question (can give up to 6 points on your final grade)

Although "gene therapy" approaches to human genetic disorders seem superficially obvious, the field has developed slowly and has been plagued with significant difficulties.

A. For the listed diseases below, put them in order for tractability and give your reasoning in each case based on the gene, the likely mutation type, and biology of the disease.

1. Huntington Disease
2. LiFraumeni (p53)
3. Beta thalassemia
4. Duschenne muscular dystrophy

B. For the most tractable one you selected in part A, outline the most plausible therapy strategy, using the current best approach. Identify the major anticipatable hazard in your design. How would test for it before trying the treatment on patients?