Bi188 Final Examination
Spring 2013

Due **Saturday, June 15th at 5:00 PM**
(as a PDF emailed to kfisher@caltech.edu, sgoh@caltech.edu, and woldb@caltech.edu).

The exam is closed-book and closed-notes.

You have 3 continuous hours to complete the exam.

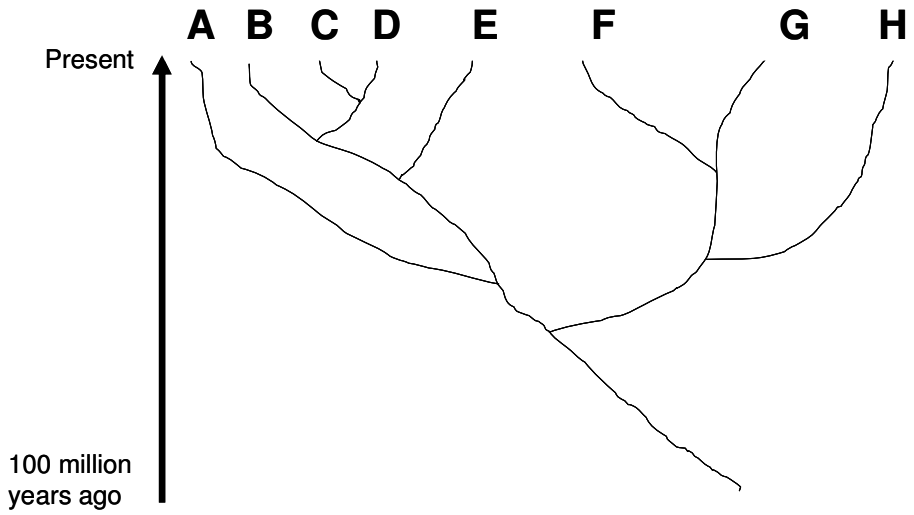
The questions are 10 points each for a total of 50 points.

You may select any 5 questions to answer. If you answer the sixth question and mark it clearly as extra credit, you can receive up to five extra credit points.

Express your answers concisely. When we ask for an experimental design, it is possible to give it in a few sentences so that the essence of the idea is there. Say what appropriate controls would be, but again, it is the appropriate idea we are asking for, not procedural detail.

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

Q1: Evolutionary genetics



Part 1. (2.5 points total, 0.5 points per section) Based on the above phylogenetic relationships, please answer the following questions: Please note that a correct answer could involve more than one species for each question.

(A) Of the following choices, Species A is most closely related to:

(i) Species F

(ii) Species G

(iii) Species H

(iv) All of the above

(B) Species A is most closely related to which species? ----------

(C) Species A is least closely related to which species? ----------

(D) The three most closely related species are: ----------

(E) The least closely related groups of species are: ----------

Part 2. (7.5 points total) You are provided with complete high quality genome sequences of 100 chimpanzees, 100 Neandertals, and 1,000 modern humans. Using your knowledge of their phylogenetic relationships and the statistical methods presented in the lecture, you decide to screen for selection that have occurred in each of these lineages. Please answer the following questions regarding your analysis:

A) Briefly define the term ‘selective sweep’, explain what it indicates in terms of selection, and describe how you would screen for selective sweeps in each lineage.
Use diagrams as needed.

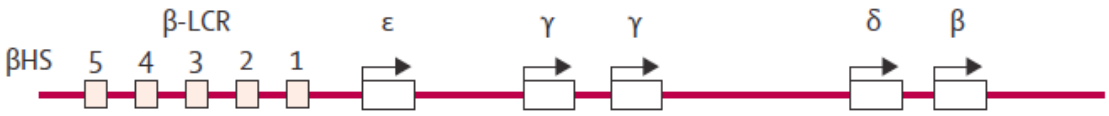
B) Briefly describe how would use the genomic information from all three species to identify candidate positive selection that occurred in (i) more recent human history (last 100,000 years) and (ii) earlier in the human lineage (around one million years ago).
Use diagrams as needed.

Q2: Globinopathies

In lecture the beta globin locus of human and mouse was discussed as a nice example of multiple principles.

A) (5 points) The human globin locus is shown schematically below. Please map onto this complex locus four different kinds of mutations: 1) a classic beta-thalassemia 2) hereditary persistence of fetal globin (HPFH) 3) a strong beta-thalassemia allele that, nevertheless, leaves ALL protein-coding sequence intact. This allele CAN be rescued by an experimental drug that 4) a very strong beta-thalassemia allele that leaves all protein coding sequences intact, is caused by sequence deletion, but could not be rescued any drug that works by inhibiting a repressor of the fetal globin.

For each of the mutant alleles above, explain why the mutation leads to the phenotype mechanistically. These explanations should be short --- a sentence or two should do it.



B) (5 points) Sketch a set of imaginary ENCODE signal tracks showing the following kinds of signals (RNA-seq; GATA-1 ChIP-Seq; BCL11a ChIP-Seq; DNAse 1 hypersensitivity) over the globin locus from above. Do this for data obtained from each of three kinds cell types from a wild-type individual. The point is for you to show something that is conceptually correct – not fine points of scaling for signals or precise locations --- show us major features of functional consequence.

1. adult erythroblasts
2. fetal erythroblasts
3. neurons

Q3: Testing for mechanisms

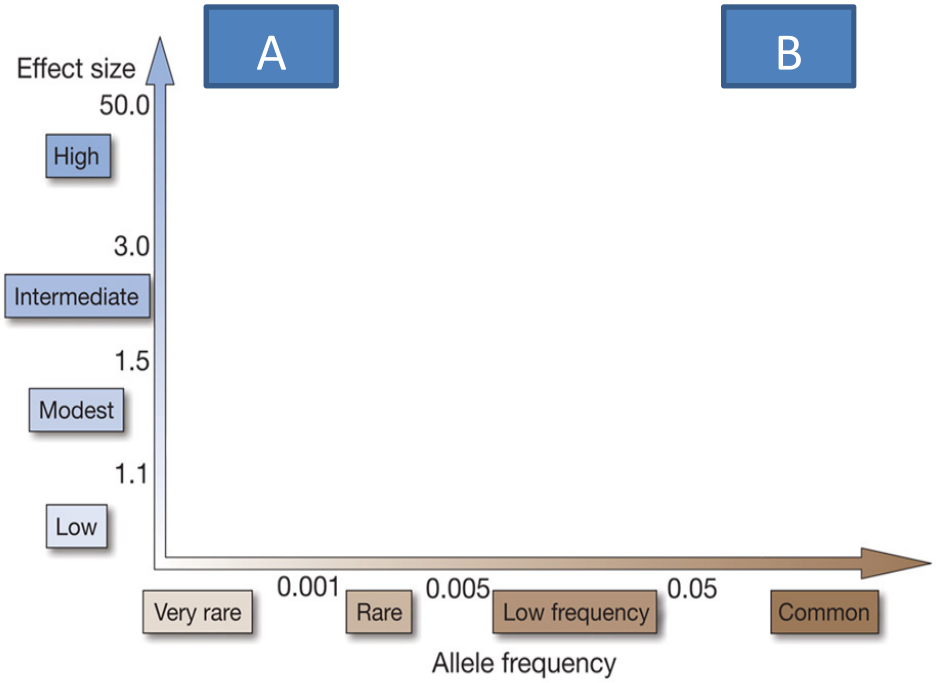
A polyA RNA-Seq study of a B-cell tumor suggests a 5-fold up-regulation of BCL2, an anti-apoptotic gene relative to non-tumor cells of the same B-cell type.

A) (2 points) Does this observation help to explain tumorgenesis in this individual? How (be brief)?

B) (4 points) If detailed sequence information covering the entire gene and a megabase in both directions around it shows no sequence difference between tumor DNA and normal cell DNA form the same individual, give two substantially different kinds of DNA level abnormalities that could explain the RNA level difference (there are at least four prominent possibilities). How would you test for each of the two and what would the key data look like? Again – be conceptual in showing what you are looking for.

C) (4 points) Start thinking about the same 5X increase in BCL2 RNA, but this time, note that an additional piece of data is from G-banding of metaphase chromosomes. There has been a translocation on another chromosome in a region containing a repressor known to bind in the promoter region of BCL2. At first you thought you would find that the repressor simply is not made because it has been mutated, but that is not what you find. Rather, RNA-Seq shows reads mapping at a higher than normal level – but only over the zinc-finger DNA binding portion of the repressor gene. Give a working hypothesis for what has happened, and show how you could mine and analyze the RNA-Seq primary data of the tumor to confirm your hypothesis.

Q4: Complex disorders



The diagram above allows one to consider potential phenotypic effects alleles of varying frequency in a given population have on human health. Based on the material discussed in class, please provide an example of a:

A) (3 points) Rare disease with very severe health effects whose genetic basis is explained by low frequency alleles in the population (Highlighted by Box A). Please state the basic inheritance pattern of this disease.

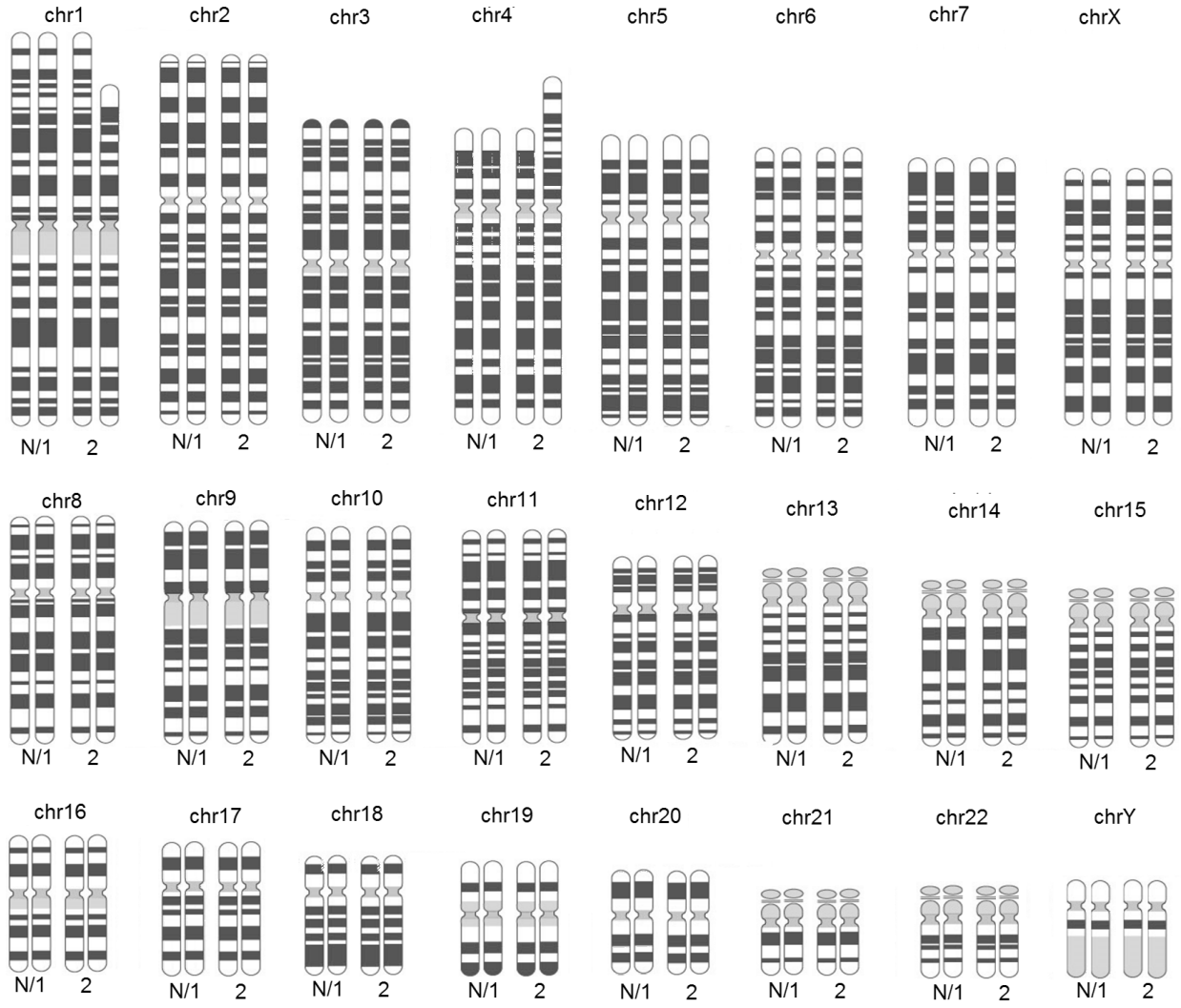
B) (3 points) Common disease with severe health effects whose genetic basis is explained by high frequency alleles in the population (Highlighted by Box B). In a few sentences, please discuss whether or not most common diseases fall into this category.

Extra credit (1 point): Provide a hypothesis based on what you know of natural selection to explain your answer above.

C) (4 points) Briefly discuss what is known about the relative genetic contribution that high frequency (common) alleles, low frequency (rare) alleles, and *de novo* mutations have on the risk of developing autism. Describe how information from GWAS and exome sequencing studies lead to your conclusions.

Q5: Chromosomal rearrangements and cancer

Your MD friend has a patient with lung cancer who is not doing well. This patient first presented 12 months ago with symptoms and primary diagnosis. At the time of first surgery, samples of his lung tumor (sample 1) were taken, as well as some healthy lung tissue (sample N). The patient has been undergoing chemotherapy, but he has developed a second site tumor and at the site of the primary tumor, some suspect tissue remains. You take a second sample tumor tissue from the primary site (sample 2) as well as a sample of his new presumptive metastasis at a distant site (sample 3). Some cells from each sample were grown briefly in culture, arrested at metaphase, and G-banded. The karyotypes for samples N and 1 are the same as each other. See below for the karyotypes of samples 1 and 2. [Ignore for this question any practical difficulties one might have with culturing of these samples]



A) (2 points) How would you find at least one mutation likely to have been causal (a driver) in this patient's initial lung cancer? In your answer, address which of the samples available you would need. How would you use data from prior large studies (TCGA, etc)

to help focus on a likely driving mutation? Do you expect one mutant driver gene in the primary tumor? Explain your expectations.

B) (2 points) Notice the difference between tumor samples 1 and 2, both of which are from the same physical location and are histologically similar. How could this have happened? Include in your answer the molecular biological explanation behind the chromosomal differences in the two samples and suggest what type of gene(s) must have been mutated in order to cause this phenomenon.

C) (2 points) You obtain RNA-Seq data from the healthy lung sample (N) as well as for tumor sample 2. You see no RNA-Seq RPKM differences between the two samples. However, there is a novel splice junction between an exon on chromosome 1 and an exon on chromosome 4. Based on this additional data, how do you think the chromosomal rearrangement in sample 2 has affected that tumor's phenotype?

D) (2 points) When you analyze sample 3, you find that the cells have undergone chromothrypsis. Predict what the karyotype for sample 3 might look like by drawing it next to the karyotypes above.

E) (2 points) You sequence the exomes of samples 2 and 3. There are no additional major protein coding sequence differences in sample 3. What other type of data would you need to acquire in order to determine what the new driving mutations are in sample 3?

Q6: Translation from genetic problem to treatment: Human models and gene therapy

You are presented with a novel childhood T-cell immunodeficiency disorder. Although the disorder is exceptionally rare, you have identified 10 affected individuals living in different regions of the world. Two patients are siblings. The other 8 patients are from completely unrelated families with no prior history of this disorder. You have obtained DNA for all affected individuals, their parents, and at least one unaffected sibling.

A) (4 points) What is your general experimental strategy for determining a focused list of candidate genes responsible for this disorder? Explain the specific leverage provided by the DNA sequence data obtained from the two affected brothers; the parents and the unaffected siblings. Briefly discuss how you would rank the most promising candidate genes.

B) (3 points) Compared with a neurodegenerative disorder, is your immunodeficiency disorder likely to be a better or worse candidate for the development of cell transplantation or gene therapy? Why? In your answer, consider the clinical trials discussed in class.

C) (3 points) You have obtained skin fibroblasts from each of these ten patients. Knowing that skin fibroblasts do not provide adequate models for this immunodeficiency disorder, you consider using these fibroblasts to generate induced pluripotent stem cells (iPSCs) as a means of studying this disease. Briefly explain the properties of iPSCs that, in theory, could make them extremely useful for investigating the genetic basis for this disease.