Bi188 Midterm Examination
Spring 2011

Due **Monday, May 2 at 5:00 PM**
(hard copy turned in to 128 Kerckhoff)
OR
Monday, May 2 at 11:59 PM
(electronic copy emailed to kfisher@caltech.edu, georgi@caltech.edu, and woldb@caltech.edu).

The exam is closed-book and closed-notes

You have 2 continuous hours to complete the exam.

The questions are 10 points total each.

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!

Question 1.

A) The human genome has ~22,000 annotated (known and confirmed) protein coding genes in the ~3 billion base pair genome. The average protein coding gene has its exons spread over ~30kb of DNA. Give the major additional kinds of functional elements (including RNA-coding and non-coding) that comprise the remaining DNA sequence that is neither an exon nor an intron (give a minimum of 3 different kinds of elements that are NOT telomeres or centromeres), and explain what their functions are. This explanation need not be lengthy - tops three sentences and a cartoon where appropriate - should do it.

B) In the beta globin locus, some mutations ameliorate the effects of deletion of adult beta globin. What is the character of these mutations and how do they work?

C) What kind of inherited mutation in the beta globin locus would leave the exons and transcriptional regulatory apparatus entirely intact for all genes in the cluster, yet nevertheless lead to a patient with same phenotype as one in which a nonsense mutation occurs very early in the protein coding region? Draw a simple cartoon to illustrate.

D) Copy number variation (CNV) is a prominent form of genetic variation in humans that was not appreciated until the human genome had been sequenced and relevant parts of multiple genomes from different individuals had been studied in detail. Why is large scale structural variation in the form of duplications or higher order tandem multiples for regions of 500 kb difficult to detect by classical DNA sequencing strategies? What experiment would you do to map CNV across the genome in individuals with developmental delay syndrome, in which brain function is broadly sub-normal? Give a cartoon sketch of what data would look like over a region with major CNV.

E) You observe unrelated human individuals with mutations that all alter (variously) the DNA sequence within a 500bp region arbitrarily called - for this problem - the USX sequence. USX is located ~1 million base pairs from the shh (sonic hedgehog) promoter. Homozygous USX mutations are associated with a phenotype of abnormal limb development. If we tell you that this phenotype is similar to the phenotype one seen after experimentally disabling the protein coding capacity of mouse shh in the developing limb (but not elsewhere), give the basic function of the region. Do you expect homozygous deletion of the protein coding domain in the mouse to have the same phenotype? If not, how and why do you expect it to differ?

Question 2.

For parts A and B of this question, you are an investigator with a collection of 100 primary ovarian tumors, normal ovarian tissue taken at the time of surgery, plus blood samples from each patient.

A) Identify two kinds of genomic and/or functional genomic data you would gather to identify candidate tumor suppressor genes (TS) (either known in other tumors or not previously discovered) whose genetic or epigenetic alteration is contributing to ovarian cancers. Explain what experiment you would do to obtain the data. You can assume you are not research budget limited. Explain how you would analyze the data to identify these candidate tumor suppressor genes. Explain how you would use information on the rate of occurrence to assess the likelihood that an observed phenomenon in your data actually contributes to ovarian cancer.

B) One ovarian tumor sample shows a rate of somatic mutation that is two orders of magnitude higher than any of the rest. What is a plausible explanation?

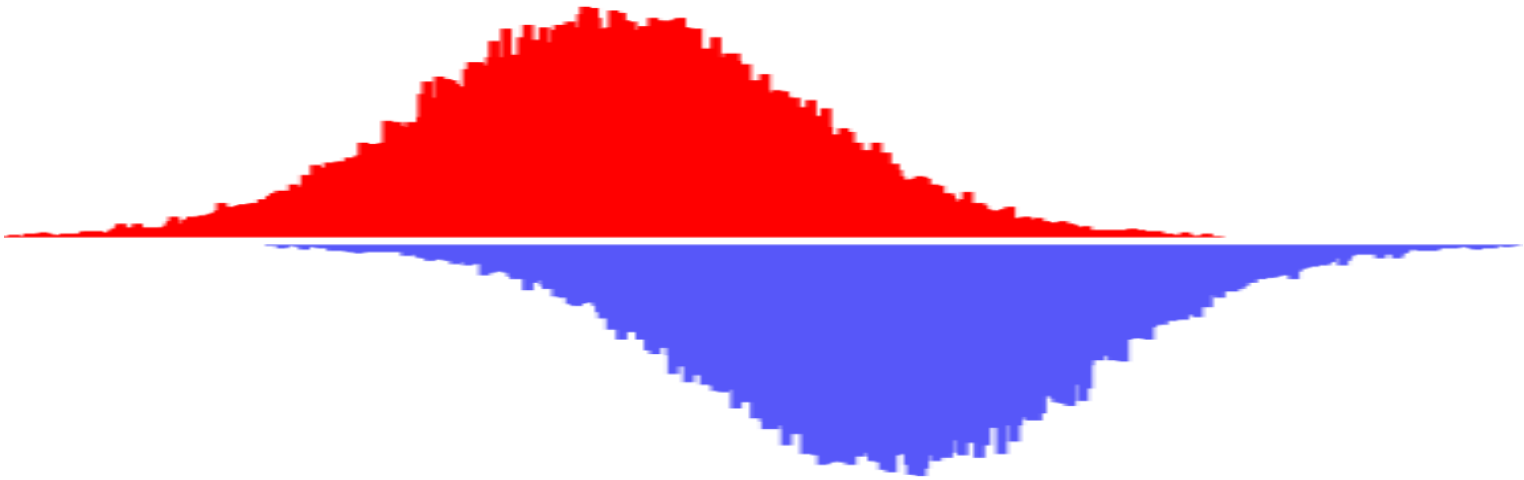
C) Genes acting as oncogenes can be activated by multiple mechanisms including point mutations, as is the case for the RAS oncogenes. What kind of additional events can also lead to oncogene activation (give two)?

D) BRCA1 and 2 were originally discovered by studying families with a high incidence of breast cancer that appeared to be inherited. A major confounding problem was that the rate of non-inherited sporadic breast cancer is very high. This confounds classical genetic mapping strategies. What additional criteria can be imposed to help focus on inherited alleles as was done in this case, and as might be done in others with a similar confound from sporadic cases? In contemporary times, on a budget that does not permit you to do entire genome sequencing, how could you identify these breast cancer tumor suppressor genes, if they were still unknown?

E) Activating mutation of the ABL kinase is causal in chronic myelogenous leukemia (CML: reminder - the Philadelphia Chromosome). A modern drug called imatinib (also known as Gleevec) was successfully designed based on knowledge of the structure of the protein encoded by the activated BCR-ABL gene. CML initially put into remission by imatinib becomes resistant to it. What is the mechanism of resistance and how could you detect/confirm it experimentally rapidly and inexpensively in any molecular biology lab? The drug is also effective in some intestinal tumors (GIST), even though the latter does not express the ABL kinase significantly. How, in principle, could the drug be effective in GIST?

Question 3.

Part 1. Below, the reference genomic sequence for 10 different regions identified as bound by transcription factor X in a ChIP-Seq experiment are shown, as well as a cumulative plus and minus strand profile for those regions. Each region is centered around its peak.



>Region1
GAGTACATCCAGCAAAAGCCGATCGGAATGGCGGCTCCACTCGACGGTGTAAGCTTACAAGGCACACAAACGACCCCACTGCACGGTAACTACGATTTTAAGCACGACAA
>Region2
TACTCGCAGTACAACGAAAGGATCCGGCAGCGACCTTGTACTCCCAGTAGGTAAGCTTACAAGAAATCAGTATATACTCGGTGACTCAAGGTCTCTAAAGGGAGGTAGGT
>Region3
AAAGGAATCTAAGGACCCTGAGATAGCTTGAAGTAGTATGGGCTCTGTAGTAAGCTTACCCGGTAACGGCTCCATCACTCGGTGGTCCTATGATGACTATATCCAATAGT
>Region4
TGGTACACCACATAAGTCTAAAACAGGCAGTCAGCACTGGCCCGCGGGTAAGCTTACAAGCCCAAACGGCATCCAGTTAGGAATATCTCTATGCCCGTGCTAGACGATTA
>Region5
AGCCCCGAACCACTAAGCCATATTAGGGTCTCCGAGGAGGGGTGCCCAGGTAAGCTTACATGTCTGGCGTGGTGAGTATTAATCACCGCTCTCTCGATTTTTCTCGTACT
>Region6
AGAATGCAGTAATGCCTGTACCCAGTCGTTTCTGCATTGCGGTCGGTGTAAGCTTACAGCCATTTTGGCCACCGCGACAACGGTGTTCGTCGCACCCACGTATTCCATGT
>Region7
GATGTCCGGTGAATTTGTTTTAATTGGGCCACAAGAGGCTGCCTTCGGCGGGTAAGCTTACGACAGCCCTGTATTCTAGTTTTAGCTGGTGTCCTCCCGGTAAGCTTACT
>Region8
GCCCCAGAGATGGAGGGGATGCCGCATACACGAGTATTAAGCGAATCACGTAAGCTTACTTTAAGGGGAGCCTGGTACCATAACAGTAACAAGGTATTTCAGCTCAACCG
>Region9
TCTTTCAATGAGTACGCCATACCGTCCGTCCCGACCACTGGCACCGGCGGCGTAAGCTTACAAAAGAAACAGATTATCACCAGCTGTAGCAGTTGGGAAATGCCCAAGAT
>Region10
GGATTGGTAAAGGACGGTGTCATTTTCTTTGCAACCTTGAGAGGAAAATGTAAGCTTACACACACTATTAGGTATAAGCGAGTCAGGCACCTTCAAGGTGCGAACGATGA

Given the data and what you know about ChIP-Seq and transcription factor biology, can you identify the recognition DNA sequence of transcription factor X? (ignore the small sample size issue). Hint: do not worry about which strand you're looking at.

Use the IUPAC convention for displaying consensus sequences shown below:

Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

Part 2. As ChIP-Seq provides a sequence readout of the ChIP assay, it makes it possible to look directly at the influence of sequence variation on transcription factor binding, for example, in the context of allelic variation when such information is available. The matched genome of the source of ChIP material in Part 1 has been sequenced as well as the genomes of its parents and allelic variants have been called and assigned as originating either from the mother or the father. The variants are shown below, as bold and underlined letter where they differ from the reference. Dashes indicate an indel.

>Region1 Maternal
GAGTACATCCAGCAAAAGCC**G**ATCGGAATGGCGGCTCCACTCGACGGTGTAA**G**CTTACAAGGCACACAAA**C**GACCCCACTGCACGGTAACTACGA----AAGCACGACAA
>Region1 Paternal
GAGTACATCCAGCAAAAGCC**C**ATCGGAATGGCGGCTCCACTCGACGGTGTAA**C**CTTACAAGGCACACAAA**A**GACCCCACTGCACGGTAACTACGA**TTTT**AAGCACGACAA

>Region2 Maternal
TACTCGCAGTACAACGAAAGGATCCGG-AGCGACCTTGTACTCCCAGTAGGTA**A**GCTTACAAGAAATCAGTATATACTCGGTGACTCAAGGTCTCTAAAGGGAGGTAGGT
>Region2 Paternal
TACTCGCAGTACAACGAAAGGATCCGG**C**AGCGACCTTGTACTCCCAGTAGGTA**T**GCTTACAAGAAATCAGTATATACTCGGTGACTCAAGGTCTCTAAAGGGAGGTAGGT

>Region3 Maternal
AAAGGA**A**TCTAAGGACCCTGAGATAGCTTGAAGTAGTATGGGCTCTGTAGTAAGCTTACCCGGTAACGGCTCCATCACTCGGTGGTCCTATGATGACTATATCCAATAGT
>Region3 Paternal
AAAGGA**G**TCTAAGGACCCTGAGATAGCTTGAAGTAGTATGGGCTCTGTAGTAAGCTTACCCGGTAACGGCTCCATCACTCGGTGGTCCTATGATGACTATATCCAATAGT

>Region4 Maternal
TGGTACACCACATAAGTCTAAAACAGGCAGTCAGCACT**G**GCCCGCGGGTAAGCTT**T**CAAGCCCAAACGGCATCCAGTTAGGAATATCTCTATGCCCGTGCTAGACGATTA
>Region4 Paternal
TGGTACACCACATAAGTCTAAAACAGGCAGTCAGCACT**T**GCCCGCGGGTAAGCTT**A**CAAGCCCAAACGGCATCCAGTTAGGAATATCTCTATGCCCGTGCTAGACGATTA

>Region5 Maternal
AGCCCCGAACCACTAAGCCATATTAGGGTCTCCGAGGAGGGGTGCCCAGGTAAGCTTACATGTCTGGCGTGGTGAGTATTAATCAC**C**GCTCTCTCGATTTTTCTCGTACT
>Region5 Paternal
AGCCCCGAACCACTAAGCCATATTAGGGTCTCCGAGGAGGGGTGCCCAGGTAAGCTTACATGTCTGGCGTGGTGAGTATTAATCAC**G**GCTCTCTCGATTTTTCTCGTACT

>Region6 Maternal
AGAATGCAGTAATGCCTGTACCCAGTCGTTTCTGCATTGCGGTCGGTGTAAGCTTACAGC**C**ATTTTGGCCACCGCGACAACGGTGTTCGTCGCACCCACGTATTCCATGT
>Region6 Paternal
AGAATGCAGTAATGCCTGTACCCAGTCGTTTCTGCATTGCGGTCGGTGTAAGCTTACAGC**A**ATTTTGGCCACCGCGACAACGGTGTTCGTCGCACCCACGTATTCCATGT

>Region7 Maternal
GATGTCCGGTGAATTTGTTTTAATTGGGCCACAAGAG**G**CTGCCTTCGGCGGGTA**T**GCTTACGACAGCCCTGTATTCTAGTTTTAGCTGGTGTCCTCCCGGTAAGCTTACT
>Region7 Paternal
GATGTCCGGTGAATTTGTTTTAATTGGGCCACAAGAG**T**CTGCCTTCGGCGGGTA**A**GCTTACGACAGCCCTGTATTCTAGTTTTAGCTGGTGTCCTCCCGGTAAGCTTACT

>Region8 Maternal
GCCCCAGAGATGGAGGGGATGCCGCATACACGAGTATTAAGCGAATCACGTAA**G**CTTACTTTAAGGGGAGCCTGGTACCATAACAGTAACAAGGTATTTCAGCTCAACCG
>Region8 Paternal
GCCCCAGAGATGGAGGGGATGCCGCATACACGAGTATTAAGCGAATCACGTAA**C**CTTACTTTAAGGGGAGCCTGGTACCATAACAGTAACAAGGTATTTCAGCTCAACCG

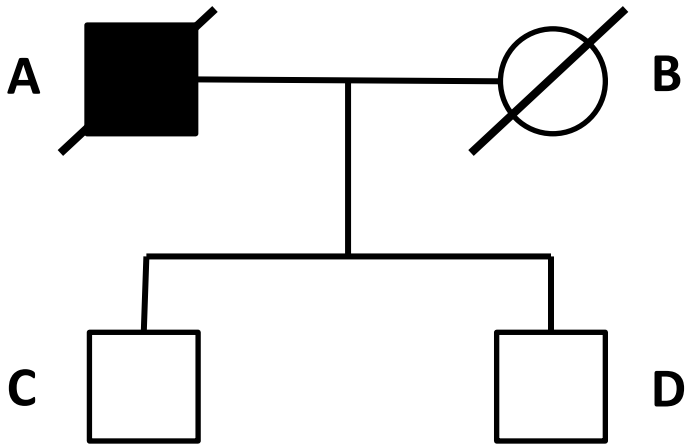
>Region9 Maternal
TCTTTCAATGAGTACGCCATACCGTCCGTCCC**GA**CCACTGGCACCGGCGGCGTAAG**C**TTACAAAAGAAACAGATTATCACCAGCTGTAGCAGTTGGGAAATGCCCAAGAT
>Region9 Paternal
TCTTTCAATGAGTACGCCATACCGTCCGTCCC**AT**CCACTGGCACCGGCGGCGTAAG**G**TTACAAAAGAAACAGATTATCACCAGCTGTAGCAGTTGGGAAATGCCCAAGAT

>Region10 Maternal
GGATTGGTAAAGGACGGTGTCATTTTCTTTGCAACCTTGAGAGGAAAATGTAAGC**T**TACACACACTATTAGGTATA**AGC**GAGTCAGGCACCTTCAAGGTGCGAACGATGA
>Region10 Paternal
GGATTGGTAAAGGACGGTGTCATTTTCTTTGCAACCTTGAGAGGAAAATGTAAGC**A**TACACACACTATTAGGTATA---GAGTCAGGCACCTTCAAGGTGCGAACGATGA

In the table below, the number of reads mapping to the parental or maternal allele are shown. Based on this information, would you make any changes in the consensus recognition sequence you derived in Part 1?

Region	Maternal Reads	Paternal Reads
1	56	52
2	103	25
3	76	80
4	34	200
5	134	131
6	85	89
7	12	47
8	123	119
9	34	39
10	267	45

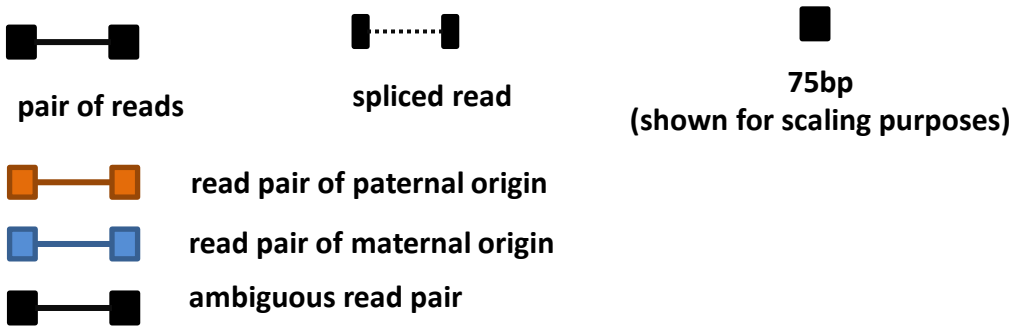
Part 3. You are studying a very rare autosomal recessive Mendelian disease that is characterized by defective function of CD4+ T cells. The specific underlying genetic cause is unknown and this is what you are trying to figure out. Unfortunately, the disease is extremely rare and all the material that is available for you to work with comes from a single family shown below (the family has a history of the disease in prior generations too).



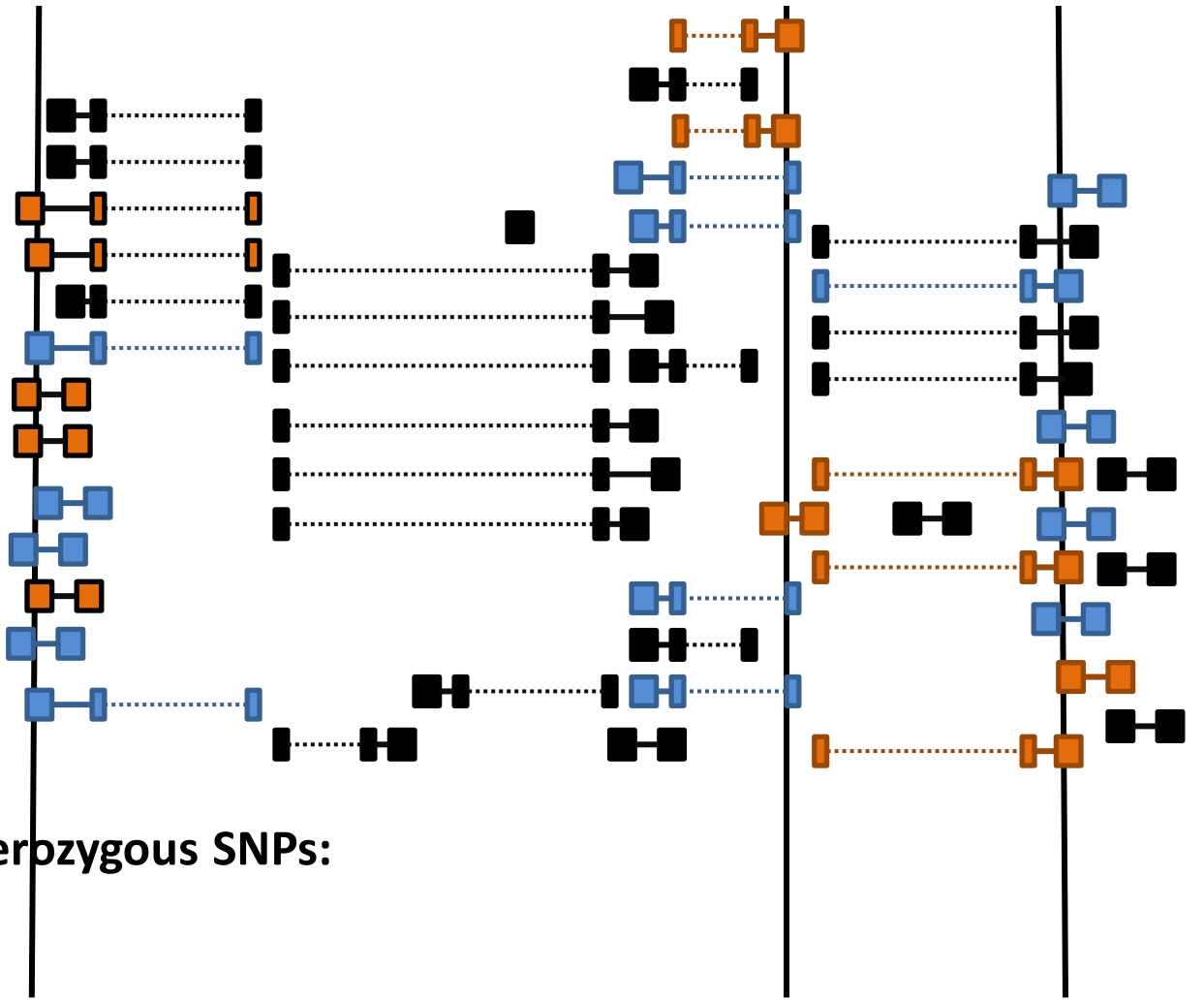
Even worse, both parents A and B are deceased so you can not isolate CD4+ cells from them. However, frozen tissues have been stored for both A and B, and you have isolated CD4+ cells from C and D. You sequence the genomes of all four individuals searching for obvious candidate mutations. Your analysis pipeline looks for previously unknown non-synonymous variants in protein coding regions. However, you do not find any of those although you see variants near or within the non-coding portions of several genes thought to be important for CD4+ cells function. You do not despair and you reason that since you have the genome sequences of both the parents and the children, you might be able to use RNA-Seq to pinpoint the variant(s) that influence the expression or function of the responsible gene on the paternal but on the maternal allele.

You do RNA-Seq on the CD4+ cells you have isolated from C and D and you align to the individuals' genome taking into account the heterozygous positions (this allows you to identify reads of maternal or paternal allele origin). One of the genomic regions that you suspect based on you previous analysis is shown below together with the allele-specific alignments and the heterozygous SNPs. Assume that CD4+ cells from both C and D give you a very similar picture. Do you think this might be the locus you are looking for and if yes, what might be the nature of the mutation and disease mechanism?

Alignment display conventions used:



Alignments:



Heterozygous SNPs:

SNP1 SNP2 SNP3