Question 1 Answers.

1A) (1pt) Asked for both RNA-coding and non-coding, so answer must include a type of RNA-coding sequence and another regulatory element such as a repressor, enhancer, or insulator. Also transposable elements, tandem repeats (may be structural). Pseudogenes accepted if they came with an argument for why they are functional. Functional elements that code for RNA include lincRNA, miRNA, rRNA, and snRNAs. (-0.5 if missing one major category or the other)

1B) (2pts) Some enhancer (or silencer) regions that can be inside or outside of the domain could allow for expression of fetal globins in later life that ameliorates or 'cancels out' the effect of the mutation of the adult ß-globin gene. Also, could have CNV for the locus and the other copies compensate for missing gene.

1C) (2pts) Splice site mutation. -1 for no illustration.

1D) (3pts) The sizes of the repeated sequences are usually much larger than the ~1kb fragments that were sequenced classically. This means that sequenced fragments would usually fall within one repeat or another and not be able to tell them apart. Only by sequencing these fragments massively high-throughput (to increase the chance you would sequence across the edge of the repeat or to informatically compare the amounts of a particular sequence) would a sequencing technology be able to detect a variation, and these methods were not available more than a few years ago.

1pt for why it's difficult to detect. 1pt for how you'd detect it (ex: array CGH/SNP microarrays). 1pt for cartoon sketch.

1E) (2pts) USX is likely to be a CRM that regulates SHH expression (1 pt). For the SHH knockout, it would likely have a more severe phenotype because SHH acts elsewhere in developing embryo besides the limbs and all of that activity would be disrupted. (1pt)

Question 2

Answer key and related discussion.

Answers are in Bold. In several places, more possible answers are lists than you were asked for. Within the answers, added explanation and information, mainly intended for clarification or for your interest, is not boldfaced.

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.

2A) 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.

Major data types discussed in class that we expected as answers are **1) Exome DNA Sequencing; 2) transcriptome sequencing; 3) whole genome DNA sequencing; 4) RNA quantification by microarrays; less expected answers that could get credit would include 5)detecting copy number variation of oncogenes by any of the CNV methods discussed; 6)global measures of microRNA representation and levels.**

For any of the genomic DNA or RNA sequence based data-types, **you would compare the normal and tumor samples for each patient to make a catalog of mutations where the tumor differs from the non-tumor tissue of the same individual.** Many of those mutations would still be along for the ride and would not be causal. You expect such "passenger" mutations to be unique to one tumor in the set – happening randomly across your modest size sample. **Those mutations that are causal in at least some of the tumors would be seen more frequently in your sample set than is expected by chance. The combination of being mutated with respect to the patient's normal tissue, and mutations of a given gene occurring with elevated frequency in the tumor set are used to argue that the gene is at least partly causal.** [Notes – In practice, candidates identified in an initial screen are often then tested by much less expensive methods in a larger tumor set to confirm that they are contribute causally to some fraction of disease and to assess what fraction that might be. Also, you expect multiple causal genes for any given tumor, so finding one is not the end of the story.] **RNA sequence could be used in a similar way to DNA sequence, though very rare RNAs (low prevalence) can**

be difficult to get strong data for; RNA data also offers the additional information on the of AMOUNT of expression, and this is relevant because oncogenes can be active because theyare overexpressed relative to their level in normal tissue. You would look for this. Tumor suppressor genes would be under-expressed relative to normal tissues, and you would look for this.

In addition, point mutations occurring in already known tumor suppressors or oncogenes would also argue to have a likely role in causation, even if they occurred in only one of your samples. The low frequency would be attributable to the relatively small sample of 100 tumors and to the diversity of gene combinations that can be involved in tumorgenesis. This answer was not required to obtain full credit, but it was good for partial credit if the rest of the answer was not complete or correct.

2B) 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?

Very high mutation rates in a tumor can arise from several causes, any one of which would get full credit as an answer: 1) The answer discussed in class was mutator genes. Their malfunction in cancer can lead to a tumor with a much higher number of mutated bases relative to somatic tissue than is typical of most tumors. [Although you were not asked to name any of these genes specifically, there are a number of genes whose mutation elevates the broader mutation rate and many of then are involved in DNA repair. Genes in this class include MSH2, MLH1, and BRCA1, the latter one being involved in some but not all ovarian tumor.] **Other correct answers to the question are: 2) high mutation accumulation due to number of mitoses in the tumor lineage and its antecedent pre-tumor cell lineage since the fertilized egg. 3) exposure of the person or the tumor to environmental agents that increase mutation rate** (like uv light for skin cancers or tobacco smoke chemicals in lung cancer). These environmental agents also include some cancer drugs, with an especially prominent case being temozolomide which is used to treat gliomas.

2C) 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)?

1) Gene amplification (ie the various Myc family members)
2) Translocations that produce a fusion oncoprotein (ie BCR-ABL in CML or PAX3/FOXO1 and PAX7/FOXO1 in rhabdomyosarcoma)
3) Mutation leading to overexpression without amplification, which could be specified to be down-regulation of a trans-acting silencer (repressor) or alteration of a cis-regulatory silencer module.
4)Epigenetic alteration of a repressing cis-regulatory element (DNA demethylation, most likely)

2D) BRCA1 was 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?

1)Cases with very early onset (patient age at diagnosis) and 2) breast cancer in males.

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?

2) Exome sequencing to identify coding mutations in families with a family history of disease in young individuals. DNA would be from archival blood or still living individuals, including postulated obligatory carriers and noncarriers. This approach depends on family history and on working in multiple unrelated families to eliminate genes with coincidentally similar allelic inheritance patterns that are irrelevant to the disease.

2E) 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?

Mutation of the drug binding site to reduce affinity for the drug

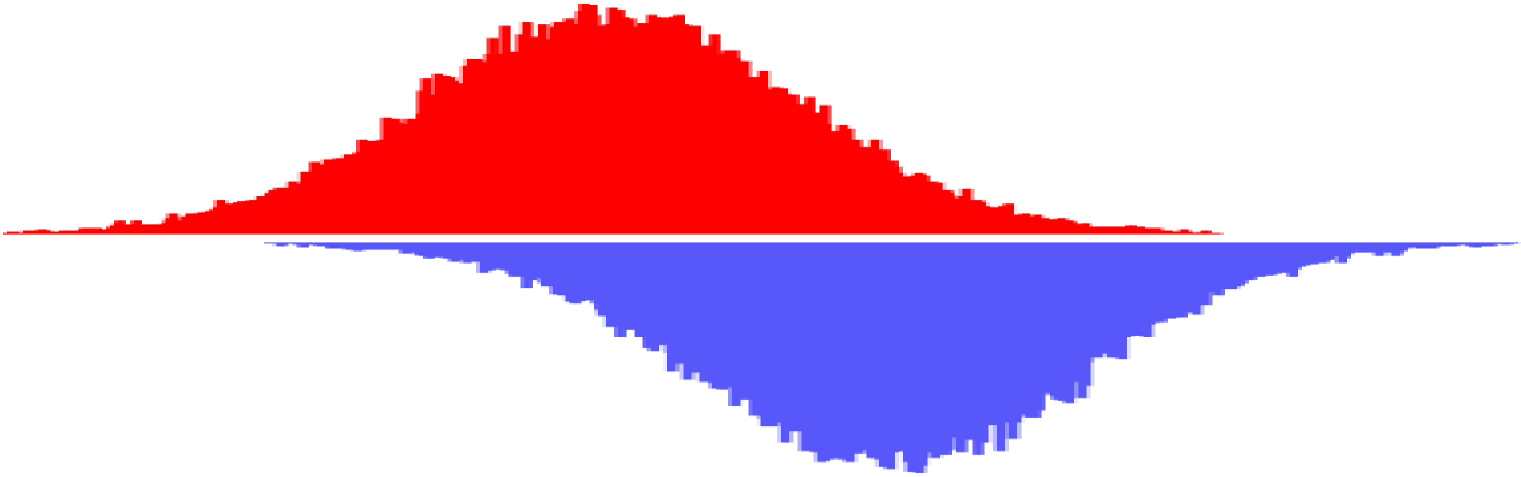
And how could you detect/confirm it experimentally rapidly and inexpensively in any molecular biology lab?

You could perform PCR on cDNA made from mRNA for BCR/ABL and sequence the PCR product to determine what changes, if any had occurred.

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?

Another kinase that is related evolutionarily to ABL (i.e. a paralogous gene) has a similar enough binding site to be preferentially bound and inhibited by the drug. [FYI, We did not ask you to name the kinase, but it is C-Kit. Cross reactivity presents a treatment for this alternate class of tumor and may also be responsible for some side effects of the drug on the class of cells from which GIST tumors are derived]

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:

Answer: GTAAGCTTAC

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

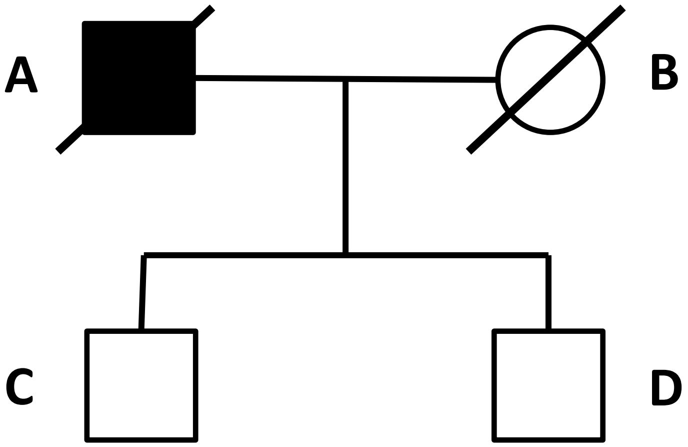
>Region10 Maternal
GGATTGGTAAAGGACGGTGTCATTTTCTTTGCAACCTTGAGAGGAAAATGTAAGC**T**TACACACACTATTAGGTATA**AGC**GAGTCAGGCACCTTCAAGGTGCGAACGATGA
>Region10 Paternal
GGATTGGTAAAGGACGGTGTCATTTTCTTTGCAACCTTGAGAGGAAAATGTAAGC**A**TACACACACTATTAGGTATA**---**GAGTCAGGCACCTTCAAGGTGCGAACGATGA

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

Answer: GTAASSTTAC

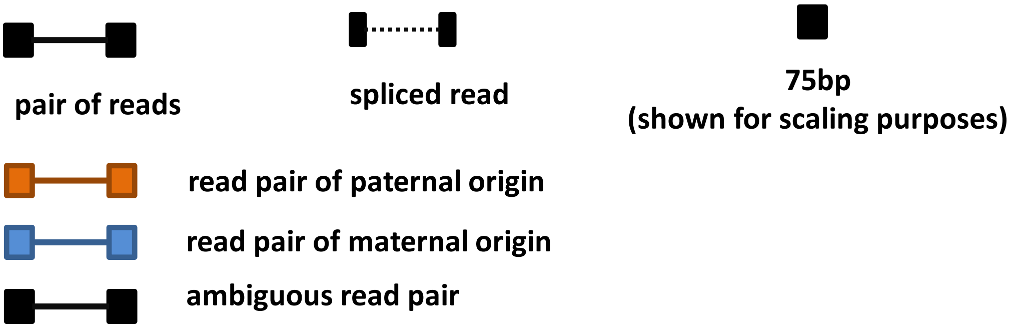
Part 3. You are studying a very rare autosomal recessive Mendelian disease that is characterized by defective function of CD4+ T cells. The specific underlying genetic cause is unknown and this is what you are trying to figure out. Unfortunately, the disease is extremely rare and all the material that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



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:



pair of reads

spliced read

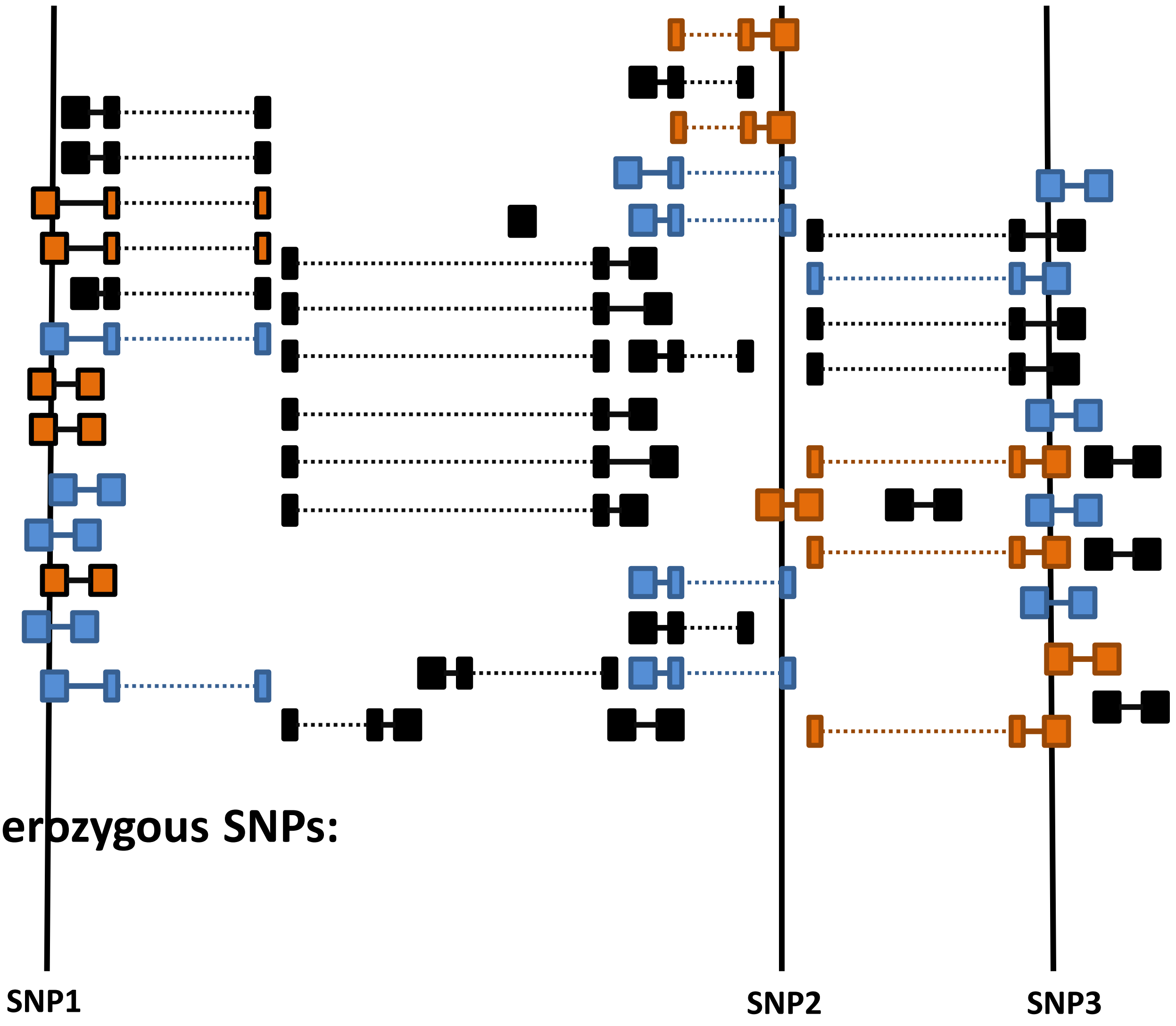
75bp
(shown for scaling purposes)

read pair of paternal origin

read pair of maternal origin

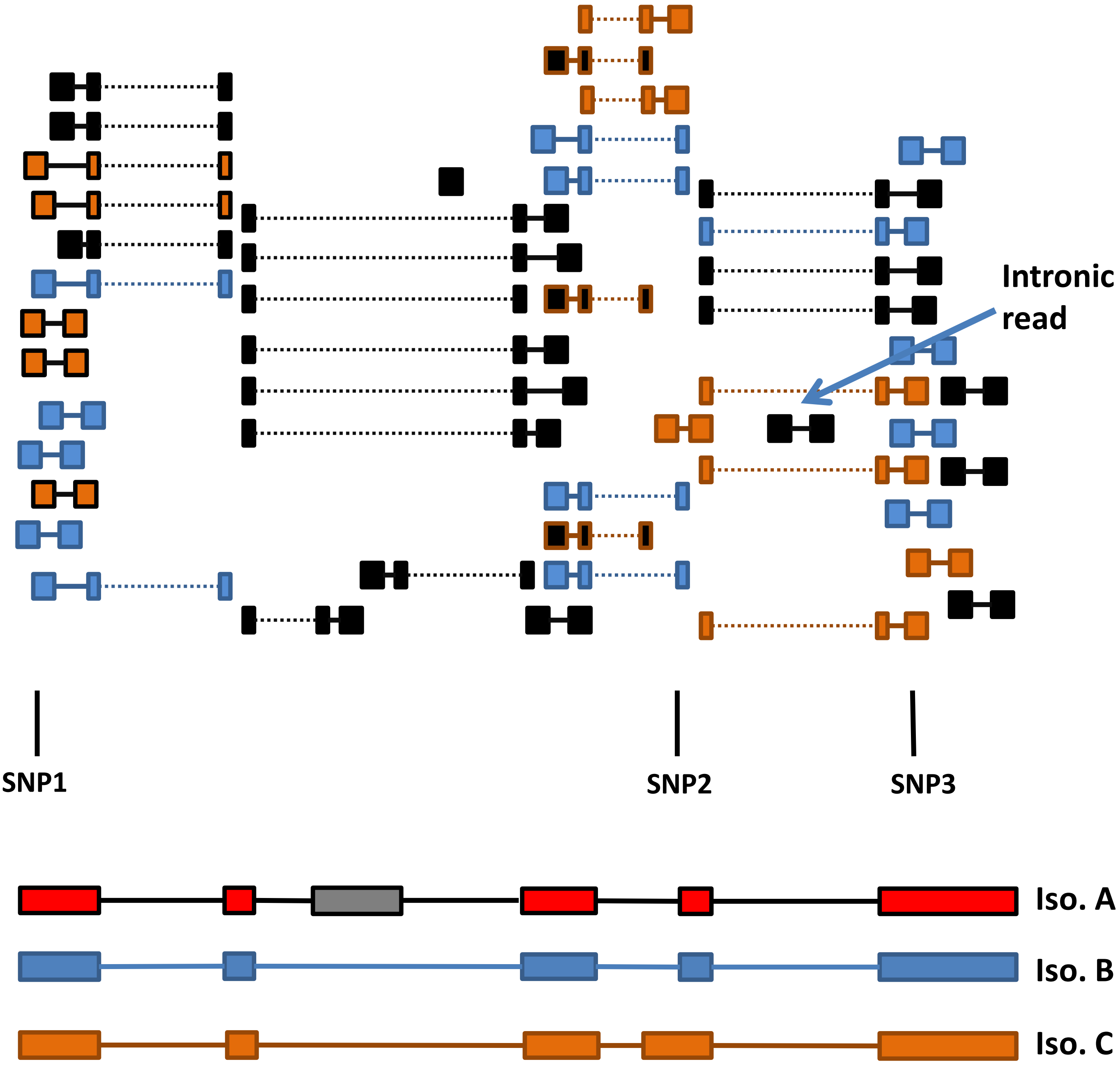
ambiguous read pair

Alignments:



Heterozygous SNPs:

Answer: Based on the splicing patterns, we can conclude that there are three isoforms being expressed. Isoform A is probably a minor one (because there is only a single spliced read supporting it compared to 6 spliced read supporting skipping that exon) and as there is no allelic difference within the exon by which it differs from the others, we can not say where it comes from (most likely it is a minor isoform from both alleles). For the other two isoforms, however, it seems like SNP #2 determines a different splicing pattern between exons 3 and 4, with exon 4 being extended in the 5' direction on the paternal chromosome. This is most likely due to SNP#2 affecting the splice site of the short exon #4. This variant may be responsible for the disease, if the resulting protein is non-functional. SNPs #1 and #3 are either in UTRs or represents synonymous or known non-synonymous changes. (it was given in the question that all no candidate non-synonymous variants were found)


Intronic read
SNP1
SNP2
SNP3
Iso. A
Iso. B
Iso. C