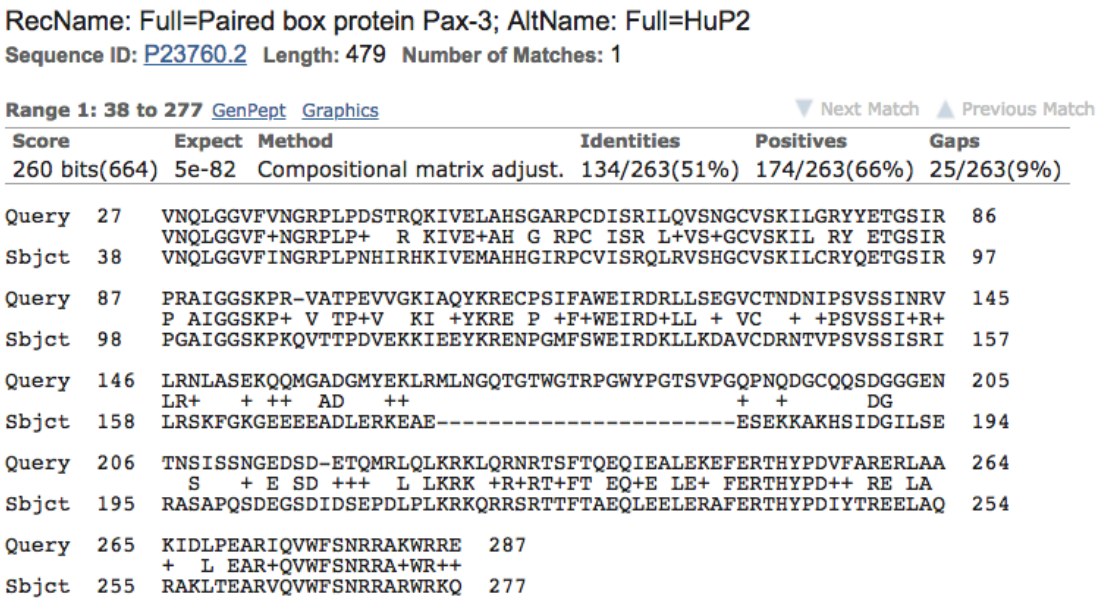
Practice problems

QUESTION 1

	1. Nucleotide mutations that do change the encoded amino acid to another amino acid are called ______ mutations.	A. Activator B. Alternative splicing C. Bootstrap analysis D. DNA E. Degenerate F. Distance matrix G. Eukaryotes H. High-complexity I. Low-complexity J. Messenger RNA (mRNA) K. Missense L. Nonsense M. Open reading frame N. Position weight matrix O. Prokaryotes P. Promoter Q. Proteins R. Qualitative S. Quantitative T. Ribosomal RNA (rRNA) U. Shine-Dalgarno sequence V. Transfer RNA (tRNA)
	2. Homology is a _____________ measure of the relationship between organisms or sequences.	
	3. One must compare ________ sequences when looking for promoters or other untranslated RNA sequences (ribosomal and transfer RNAs, for example).	
	4. ________________regions will often align well with one another, but the alignment is not due to homology but by chance.	
	5. The genetic code is ______________means that most amino acids can be specified by more than one codon.	
	6. __________ is the main reason behind the fact these genomes have a much small number of genes when compared to the number of proteins.	
	7. The major difference between eukaryotes and prokaryotes in terms of their transcription and translation processes is that the ___________ mRNA transcripts are substantially modified before translation.	
	8. The most important core ________ sequence in genes transcribed by RNA polymerase II is called the TATA box. This sequence is characterized by TATA sequence motif.	
	9. The ________________is the physical link between the mRNA and the growing protein chain.	
	10. ________________is designed to estimate the robustness of the constructed phylogenetic tree. It is based on repeating the tree construction for different samplings of the same dataset.	

QUESTION 2

Below is one of the BLAST hits. For each ALIGNMENT, the QUERY sequence ("Query") is shown at the top and the hit ("Sbjct") underneath it, with the position of the amino acids indicated on the right and left of the alignment.



RecName: Full=Paired box protein Pax-3; AltName: Full=HuP2

Sequence ID: P23760.2 **Length:** 479 **Number of Matches:** 1

Range 1: 38 to 277 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
260 bits(664)	5e-82	Compositional matrix adjust.	134/263(51%)	174/263(66%)	25/263(9%)

```
Query  27   VNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIR  86
            VNQLGGVF+NGRPLP+  R KIVE+AH G RPC ISR L+VS+GCVSKIL RY ETGSIR
Sbjct  38   VNQLGGVFINGRPLPNHIRHKIVEMAHHGIRPCVISRQLRVSHGCVSKILCRYQETGSIR  97

Query  87   PRAIGGSKPR-VATPEVVGKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRV  145
            P AIGGSKP+ V TP+V  KI +YKRE P +F+WEIRD+LL + VC  + +PSVSSI+R+
Sbjct  98   PGAIGGSKPKQVTTPDVEKKIEEYKRENPGMFSWEIRDKLLKDAVCDRNTVPSVSSISRI  157

Query  146  LRNLASEKQQMGADGMYEKLRMLNGQTGTWGTRPGWYPGTSVPGQPNQDGCQQSDGGGEN  205
            LR+   + ++  AD   ++                         +  +      DG
Sbjct  158  LRSKFGKGEEEEADLERKEAE-----------------------ESEKKAKHSIDGILSE  194

Query  206  TNSISSNGEDSD-ETQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAA  264
              S   + E SD +++  L LKRK +R+RT+FT EQ+E LE+ FERTHYPD++ RE LA
Sbjct  195  RASAPQSDEGSDIDSEPDLPLKRKQRRSRTTFTAEQLEELERAFERTHYPDIYTREELAQ  254

Query  265  KIDLPEARIQVWFSNRRAKWRRE  287
            +  L EAR+QVWFSNRRA+WR++
Sbjct  255  RAKLTEARVQVWFSNRRARWRKQ  277
```

Answer the following questions:

1) In the Sbjct sequence between 158 and 194, what do the stretches "---" represent?

____

2) What is the name of the gene that this BLAST hit returns ________________________

3) What is the degree of similarity between the query and the hit? __________________

4) What is the statistical probability that the similarity between the query and the hit occurs only by chance?

____

5) What does the '+' signs between the query and hit stand for?

____

QUESTION 3

1) The nucleotide sequence of one DNA strand of a double helix is given. Write the complementary sequence found on the other strand. Label the 5’ and 3’ ends of the molecule.

```
5' --- GACAGTCATGGCTTTTGA --- 3'
```

2) Suppose that the DNA molecule above is transcribed and the lower strand is used as the template strand. What is the RNA sequence obtained from the transcription? Label the 5’ and 3’ ends of the molecule.

3) How many possible reading frames are there for the following sequence (do not need to list out all the reading frames)?

```
5' --- GCACTAGTCAAGGCTTTTGAC --- 3'
```

4) Complete the following table. Label 5’ and 3’ ends of DNA and RNA, and the amino and carboxyl ends of proteins. Assume that
- the reading is from left to right
- the columns represent transcriptional and translational alignments

	C													**DNA double helix**
				T	C	A								
		C	A											**mRNA transcribed**
										G	C	A		**Appropriate tRNA anticodon**
							Trp (W)							**Amino acids incorporated into protein**

5) True or False. Write **T** for True, **F** for False.

a) The key assumption made when constructing a phylogenetic tree from a set of sequences is that they are all derived from a single ancestral sequence. ____

b) When comparing data from two distantly related species, the rapidly changing regions will show almost uniform dissimilarity, but the more conserved regions will convey useful information for the construction of phylogenetic trees. ____

c) Cladograms assumes a constant rate of mutation along all branches, known as the molecular clock assumption. ____

d) Using a protein sequence, one can perform BLAST search using the blastx algorithm. ____

e) Because the two strands of DNA are complementary, the mRNA of a given gene can be synthesized using either strand as a template. ____

QUESTION 4

```
LOCUS       HSU14680              5711 bp    mRNA    linear   PRI 10-JUN-2002
DEFINITION  Homo sapiens breast and ovarian cancer susceptibility (BRCA1)
            mRNA,complete cds.
ACCESSION   U14680
FEATURES             Location/Qualifiers
     source          1..5711
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /chromosome="17"
                     /map="17q21; spans D17S855"
     gene            1..5711
                     /gene="BRCA1"
     exon            1..100
                     /gene="BRCA1"
                     /number=1
     exon            101..199
                     /gene="BRCA1"
                     /number=2
     CDS             120..5711
                     /gene="BRCA1"
                     /codon_start=1
                     /product="breast and ovarian cancer susceptibility"
                     /protein_id="AAA73985.1"
     exon            200..253
                     /gene="BRCA1"
                     /number=3
ORIGIN
        1 agctcgctga gacttcctgg accccgcacc aggctgtggg gtttctcaga taactgggcc
       61 cctgcgctca ggaggccttc accctctgct ctgggtaaag ttcattggaa cagaaagaaa
      121 tggatttatc tgctcttcgc gttgaagaag tacaaaatgt cattaatgct atgcagaaaa
      181 tcttagagtg tcccatctgt ctggagttga tcaaggaacc tgtctccaca aagtgtgacc
      241 acatattttg caaattttgc atgctgaaac ttctcaacca gaagaaaggg ccttcacagt
      ...
```

1) What are the first 3 amino acids of the protein encoded by this gene? ____ ____ ____

2) What is the length of the protein encoded by this gene? ___________

3) What are the start and end positions of the 5'-UTR of this gene? Start ________ End ______

4) For each of the following three mutations, mark the mutation on the sequence and explain what consequences on the protein it might have, if possible, include the amino acid change in your answer. Clearly give details as was done in class for similar problems.

   a) Exon 1, 57 G → A

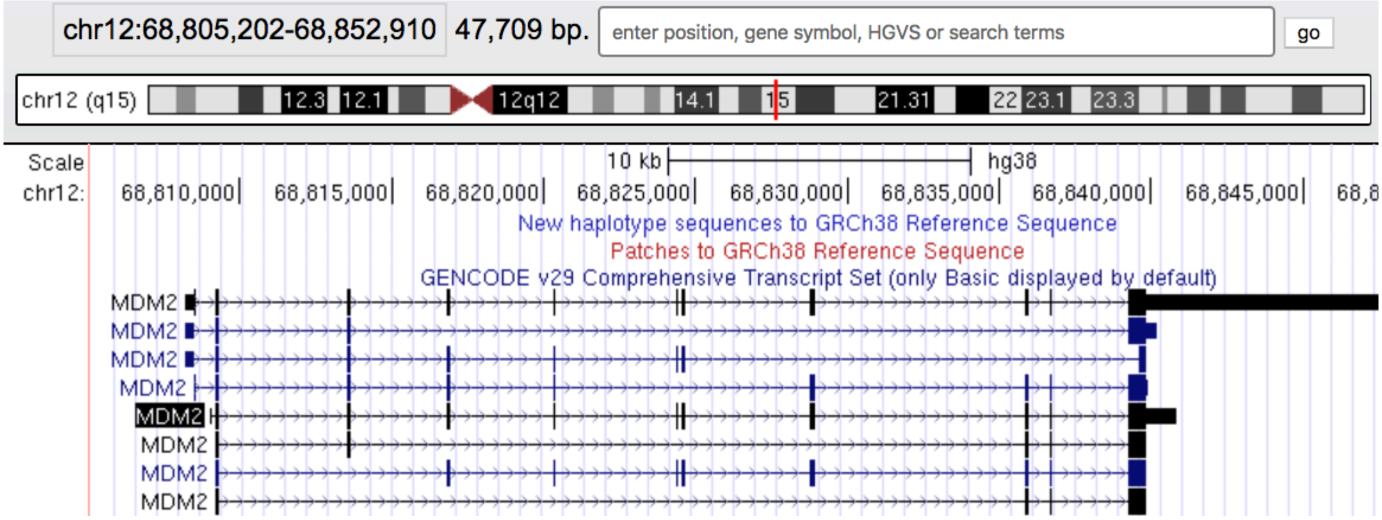
   b) Exon 2, 122 G → A

   c) Exon 3, 200 del T

QUESTION 5
Fill in the blanks:

1. The height of the sequence logo represents the ________________________________.

2. Answer the following questions using the screenshot below:


chr12:68,805,202-68,852,910 47,709 bp.

enter position, gene symbol, HGVS or search terms
go
chr12 (q15)
12.3
12.1
12q12
14.1
15
21.31
22
23.1
23.3
Scale
10 kb
hg38
chr12:
68,810,000
68,815,000
68,820,000
68,825,000
68,830,000
68,835,000
68,840,000
68,845,000
68,8
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
MDM2
MDM2
MDM2
MDM2
MDM2
MDM2
MDM2
MDM2

a) What strand is this gene locate? ____

b) How many exons are in the MDM2 transcript that is highlighted above? __________

c) What gene is being displayed here? ____

d) Circle the 3'-UTR regions on the MDM2 transcript that is highlighted above. Be specific.

e) What are the genomic regions that contain →→→→ symbol? ____________

f) Put a 'X' on third exon on the highlighted MDM2 transcript above.

g) Which the tool generates this display? ______________________________________

QUESTION 6

1. Consider the following lines from a fastq file generated by Illumina, a next-generation sequencing platform:

```
@SRR5680996.31734339 D64TDFP1:315:C7T1RACXX:2:2316:17363:71358/1
AGAGCAACACCTTGTGCCTCCAAGAAAGTATTAGTCTCCCTGAGGACTCTT
+
@@@DDD?BD??DFHCBGHEFGCBBGIGD1?EGIGHG@DDGIEG>DGGIFII
```

Explain the content of these 4 lines above:

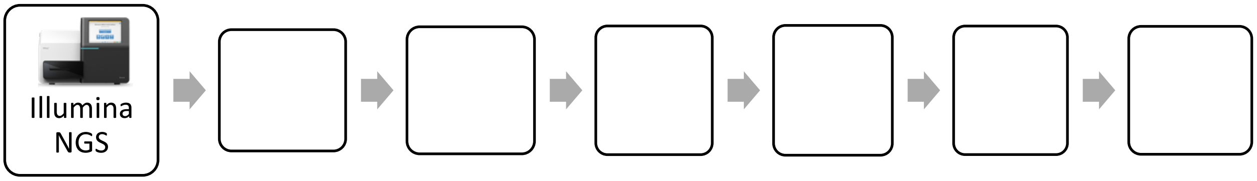
Line 1: ____

Line 2: ____

Line 3: ____

Line 4: ____

2. Fill in the blanks for the NGS workflow typically employed for variant discovery or genotyping. Select from the choices and place the letter representing the answer in each box below:



   A. Perform variant calling
   B. Trimming low base quality and sequencing adapters from the ends of the reads
   C. Mark duplicate reads in the aligned reads
   D. Align sequencing reads to the reference genome
   E. Read Quality Control using FastQC
   F. Recalibrate base quality scores

3. What are the 7 steps in preparing the DNA extracted from a biological sample for Illumina NGS?